Journal of Mathematics and Physics

(Founded by C. L. E. Moore)

Volume XXXII

1953 – 54

Board of Editors

E. REISSNER, *Managing Editor*

H. FESHBACH
P. FRANKLIN
E. A. GUILLEMIN
Z. KOPAL
C. C. LIN
P. M. MORSE
D. J. STRUIK
N. WIENER

The Technology Press
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

QA1
J87
v.32

CONTENTS

Page

Synthesis of 2n-Poles by Networks Containing the Minimum Number of Elements
By B. D. H. Tellegen 1

Spinor Theory of Four-Terminal Networks
By W. T. Payne 19

Secondary Flow Within a Bend
By Hans P. Eichenberger 34

A Heat Conduction Problem and the Product of Two Error Functions
By Earl D. Rainville 43

On a Class of Stochastic Operators
By L. A. Zadeh 48

Upper and Lower Bounds for Torsional Rigidity
By H. F. Weinberger 54

An Abridged Block Method for the Solution of the Dirichlet Problem for the Laplace Difference Equation
By Charles Saltzer 63

Note on the Fitting of Polynomials to Equally-Spaced Observations
By P. G. Guest 68

On the Transonic Flow Past a Finite Wedge
By L. Trilling and K. Walker, Jr. 72

On a Stability Criterion for Partial Difference Equations
By Stephen H. Crandall 80

On the Flow of an Incompressible Viscous Fluid Past a Flat Plate at Moderate Reynolds Numbers
By Y. H. Kuo 83

Diffraction and Reflection of Weak Shocks by Structures
By Lu Ting 102

Bounds for Solutions of Boundary Value Problems in Elasticity
By K. Washizu 117

On a Variational Theorem for Finite Elastic Deformations
By Eric Reissner 129

A Mathematical Basis for an Error Analysis of Differential Analyzers
By K. S. Miller and F. J. Murray 136

Note on S. N. Lin's Method of Factoring Polynomials
By F. B. Hildebrand 164

A New Integration Procedure
By M. Lotkin 171

The Vector Method of Solving Simultaneous Linear Equations
By Everett W. Purcell 180

On the Solution of the Differential Equation Occurring in the Problem of Heat Convection in Laminar Flow Through a Tube
By Milton Abramowitz 184

Evaluation of the Integral $\int_0^\infty e^{-u^2-x/u}\,du$
By Milton Abramowitz 188

Note on the Zeros of $(dP^1_{m_i}(x)/dx)|_{x=0}$
By K. M. Siegel, J. W. Crispin, R. E. Kleinman and H. E. Hunter 193

Radix Table for Obtaining Hyperbolic and Inverse Hyperbolic Functions to Many Places
By Herbert E. Salzer 197

A Note on the Network Postulates
By Maynard G. Arsove 203

A 39673

Page

A Note on Contact Transformations
By H. D. Block 207
"Monte Carlo" Methods for the Iteration of Linear Operators
By J. H. Curtiss 209
A General Solution of the Two-Frequency Modulation Product Problem. I.
By R. L. Sternberg and H. Kaufman 233
On Richardson's Method for Solving Linear Systems with Positive Definite Matrices
By David Young 243
A System of Non-Linear Partial Differential Equations Arising in Heat Transfer
By S. Katz and A. M. Peiser 256
Recurrence Relations for Prolate Spheriodal Wave Functions
By Imanuel Marx 269
Dynamic Analysis of the Buckling of Laterally Loaded Flat Arches
By N. J. Hoff and Victor G. Bruce 276
On the Numerical Solution of Volterra Integral Equations
By Carl Wagner 289
A Note on Sets of Matrices Simultaneously Reducible to the Triangular Skeleton
By J. C. Eaves 302
Note on an Equi-Momental Complex of a Rigid Body
By Hari Das Bagchi 307
Boundary Layer Control by Distributed Surface Suction or Injection. Bi-Parametric General Solution
By T. P. Torda 312
A Note on the Paper "On a Class of Iterative Procedures for Solving Normal Systems of Ordinary Differential Equations"
By J. Warga 315

SYNTHESIS OF $2n$-POLES BY NETWORKS CONTAINING THE MINIMUM NUMBER OF ELEMENTS

By B. D. H. Tellegen

1. Statement of the problem. The synthesis of passive two-poles was accomplished by Brune.[1] Brune's networks contain the minimum number of reactances (inductances and capacitances), equal to the highest power p of the frequency parameter λ in numerator and denominator of the impedance function of the two-pole. Brune's networks also contain the minimum number of elements (inductances, capacitances, resistances, and ideal transformers), equal to the number $2p + 1$ of independent parameters defining the impedance function. Darlington,[2] Cauer,[3] and Cocci[4] have shown that it is possible to realize any impedance function by a network containing not more than one resistance. These networks contain, however, more than the minimum number of reactances, and more than the minimum number of elements.

The synthesis of passive four-poles was accomplished by Gewertz.[5] As Leroy[6] pointed out, Gewertz' method is rather to be regarded as an extension of the method of Darlington e.a. than as an extension of the method of Brune. For by using Darlington's results the method of Gewertz leads to networks containing not more than two resistances, and, as Leroy[6] has shown, this is the minimum number of resistances necessary to realize a general four-pole. Also, in a certain sense, it can be said that Gewertz' networks contain more than the minimum number of reactances and more than the minimum number of elements. These statements get a definite meaning only after introducing the concept of *order of a four-pole.*[7] The order of a four-pole is defined as the order of the differential equation of the free oscillations of the system formed by connecting the terminal pairs of the four-pole to the terminal pairs of a four-pole containing no reactances. The minimum number of reactances is then equal to the order p of the four-pole, and the minimum number of elements, being equal to the number of independent parameters defining a general four-pole of order p, is equal to $3p + 3$, as will be proved in section 4.

The method of Gewertz was, in several ways, extended to passive $2n$-poles by Bayard,[8] Oono,[9] Leroy,[10] and Belevitch.[11] They derived networks con-

[1] O. Brune, J. Math. Phys. **10**, 191, 1931.
[2] S. Darlington, J. Math. Phys. **18**, 257, 1939.
[3] W. Cauer, Theorie der linearen Wechselstromschaltungen, 1941, p. 437.
[4] G. Cocci, Alta Frequenza, **9**, 685, 1940.
[5] C. Gewertz, J. Math. Phys. **12**, 1, 1933.
[6] R. Leroy, Câbles & Transmission, Paris, **3**, 141, 1949.
[7] B. D. H. Tellegen, Philips Res. Rep. **1**, 169, 1946.
[8] M. Bayard, Bull. Soc. franç. Élect. **9**, 497, 1949; Câbles & Transmission, Paris, **4**, 281, 1950.
[9] Y. Oono, J. Math. Phys. **29**, 13, 1950.
[10] R. Leroy, Câbles & Transmission, Paris, **4**, 234, 1950.
[11] V. Belevitch, Ann. Télécommun. **6**, 302, 1951.

MATHEMATICS GRADUATE LIBRARY

taining the minimum number of resistances, generally n, but more than the minimum number of reactances and more than the minimum number of elements.

The object of this paper is to extend the method of Brune to the synthesis of passive $2n$-poles, and thereby to arrive at networks containing the minimum number of reactances and the minimum number of elements.[12] Section 2 contains some preliminary considerations. The synthesis procedure is described in section 3, while most of the proofs are given in later sections.*

2. Preliminary considerations. To synthesize a $2n$-pole we start from its impedance matrix

$$\| Z_{ik} \| ; \qquad Z_{ik} = Z_{ki}, Z_{ik} = R_{ik} + jX_{ik} ; \qquad i, k = 1, 2, \cdots, n.$$

The Z_{ik} are rational functions of the frequency parameter λ. It is known that for a passive $2n$-pole this matrix is a positive real matrix,[13] i.e. the quadratic form associated with the matrix is a positive real function:

$$\sum_{i,k=1}^{n} Z_{ik} x_i x_k \tag{1}$$

is a positive real function for all real values of the x's. A rational function $f(\lambda)$ is said to be a positive real function if

(a) $f(\lambda)$ is real for λ real,

(b) Re $f(\lambda) \geqq 0$ for Re $\lambda \geqq 0$.

Brune[14] has shown that a necessary and sufficient condition for a rational function $f(\lambda)$ to be a positive real function is that

(a) the coefficients are real,

(b) Re $f(\lambda) \geqq 0$ on the imaginary axis of the λ-plane,

(c) $f(\lambda)$ has no poles in the right half of the λ-plane,

(d) any poles on the imaginary axis are simple and have positive real residues.

By applying this to (1) we find that a necessary and sufficient condition for a matrix $\| Z_{ik} \|$ whose elements are rational functions of λ to be a positive real matrix is that

(a) the coefficients of the Z_{ik} are real,

[12] Oono, reference 9, at the end of his paper, mentions that in another paper he has given networks containing no superfluous elements. As this paper was written in Japanese, I have not been able to read it.

[13] C. Gewertz, reference 5, pp. 53, 85. E. A. Guillemin, Communication networks, Vol. II, 1935, p. 217. W. Cauer, reference 3, pp. 127, 136.

[14] Reference 1, p. 204.

* Note added in proof. After the writing of the paper there appeared an extensive study by B. McMillan, Bell Syst. tech. J. **31**, pp. 217, 541, 1952, dealing with the synthesis of $2n$-poles. The procedure of McMillan has much in common with mine; his notion of "degree" of a $2n$-pole corresponds to my notion of "order." During a visit to the Bell Telephone Laboratories in 1952 Mr. Dietzold showed me an English translation of the paper by Oono mentioned in note 12, which proved to me that Oono extended the Brune procedure to the synthesis of $2n$-poles in a way different from mine. At the symposium on modern network synthesis in New York, April 1952, Mr. Bayard surveyed the work on the synthesis of $2n$-poles (Proc. of the Symp. p. 66; also Ann. Telecommun. **7**, 517, 1952).

(b) $| R_{ik} |$ and all principal minors are $\geqq 0$ on the imaginary axis,
(c) the Z_{ik} have no poles in the right half plane,
(d) any common poles of the Z_{ik} on the imaginary axis are simple and have real residues r_{ik} for which $| r_{ik} |$ and all principal minors are $\geqq 0$.

The inverse matrix of a positive real matrix is also positive real.[15] The inverse matrix of the impedance matrix $\| Z_{ik} \|$ is the admittance matrix $\| Y_{ik} \|$; so if we know that the impedance matrix is positive real, the same holds for the admittance matrix, and vice versa.

When one or more of the elements of a matrix have a pole at a certain value of λ, we say that the matrix has a *pole* at that value of λ. When the determinant of a matrix is zero at a certain value of λ, we say that the matrix has a *zero* at that value of λ. At a zero of a matrix the inverse matrix has a pole. The rank of the matrix of the residues of a pole we call the *rank of the pole*. For a zero the rank of the corresponding pole of the inverse matrix we call the *rank of the zero*. When the determinant of a matrix is zero for all values of λ, the matrix is said to be *singular*.

A pole of an impedance matrix at $\lambda = 0$ can be split off by writing

$$\| Z_{ik} \| = \| r_{ik}/\lambda \| + \| Z'_{ik} \| , \tag{2}$$

where the r_{ik} are the residues of the pole. From the afore-mentioned properties of positive real matrices it follows that the matrices on the right-hand side of (2) are both positive real. The first of these matrices can be realized by a $2n$-pole composed of capacitances and ideal transformers.[16] The minimum number of capacitances is equal to the rank of the pole. If the rank is one, the $2n$-pole can consist of one capacitance and $n - 1$ ideal transformers. Fig. 1 illustrates the splitting off of such a pole by means of a series $2n$-pole, for $n = 3$. Fig. 2 illustrates the splitting off of a pole of rank one at $\lambda = 0$ of an admittance matrix by splitting off a parallel $2n$-pole, also for $n = 3$. The splitting off of poles at $\lambda = \infty$ or at two conjugate imaginary values of λ can be illustrated by fig. 1 after changing the capacitance into an inductance resp. into a capacitance and an inductance in parallel, and by fig. 2 after changing the inductance into a capacitance resp. into an inductance and a capacitance in series. Zeros of a matrix on the imaginary axis can be split off by splitting off the corresponding poles of the inverse matrix.

The order of the differential equation of the free oscillations of the system formed by connecting the terminal pairs of a $2n$-pole to the terminal pairs of a $2n$-pole containing no reactances we call *the order of the $2n$-pole*. For $n = 1$ the order is equal to the highest power of λ in numerator and denominator of the impedance function. The order of a $2n$-pole cannot be larger than the number of reactances contained in the $2n$-pole network. Thus the number of reactances necessary to realize a $2n$-pole is equal to its order.

[15] C. Gewertz, reference 5, p. 59. W. Cauer, reference 3, p. 133. R. Leroy, reference 10, Théorème IV.

[16] W. Cauer, reference 3, p. 190.

In the next section we give a synthesis procedure for general $2n$-poles of a given order. We suppose the impedance matrix to have no poles or zeros on the imaginary axis, for these can first be split off by the procedures described above. Our synthesis procedure demonstrates that every positive real matrix can be realized as the impedance matrix of a passive $2n$-pole, as was already shown by Bayard and Oono.[17] As we can also start from the admittance matrix and use analogous procedures, every positive real matrix can also be realized as the admittance matrix of a passive $2n$-pole.

3. The synthesis procedure. The synthesis procedure consists in developing a zero of the impedance matrix on the imaginary axis of the λ-plane and splitting off the zero by splitting off the corresponding pole of the inverse matrix. To de-

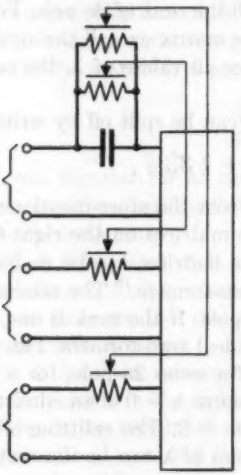


FIG. 1

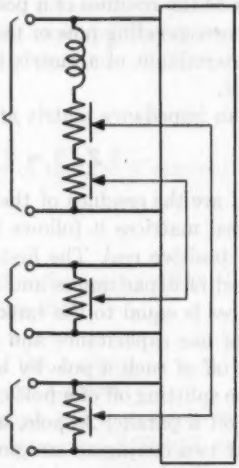


FIG. 2

FIG. 1. The splitting off of a pole of rank one at $\lambda = 0$ of an impedance matrix.
FIG. 2. The splitting off of a pole of rank one at $\lambda = 0$ of an admittance matrix.

velop such a zero we begin by splitting off a resistance R in series with the first terminal pair. The impedance matrix will remain positive as long as

$$\begin{vmatrix} R_{11} - R & R_{12} & \cdots & R_{1n} \\ R_{12} & R_{22} & \cdots & R_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ R_{1n} & R_{2n} & \cdots & R_{nn} \end{vmatrix} \geqq 0$$

on the imaginary axis. If we put $| R_{ik} | = \Delta$ and denote the minor of the element R_{11} by Δ_{11}, we can write the condition as

$$R \leqq \Delta/\Delta_{11} .$$

[17] References 8 and 9.

We determine the minimum of Δ/Δ_{11} on the imaginary axis and make R equal to this minimum. The minimum can lie at zero frequency, at infinite frequency, or at a finite value of the frequency. We shall consider these cases separately. The first two cases give no trouble; the third case is the difficult one.

(a) *The minimum lies at zero frequency.* In this case the impedance matrix, after splitting off R, has a zero at $\lambda = 0$. Thus the admittance matrix has a pole at $\lambda = 0$. The pole is generally of rank one (section 5) and can thus generally be split off by a parallel $2n$-pole consisting of an inductance and $n - 1$ ideal transformers (fig. 2). By this procedure the order of the $2n$-pole is lowered by one (section 6).

(b) *The minimum lies at infinite frequency.* The corresponding zero at $\lambda = \infty$ can, analogous to (a), generally be split off by a parallel $2n$-pole consisting of a capacitance and $n - 1$ ideal transformers. This also lowers the order of the $2n$-pole by one.

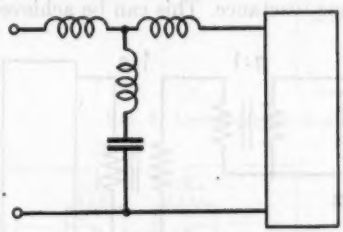


FIG. 3

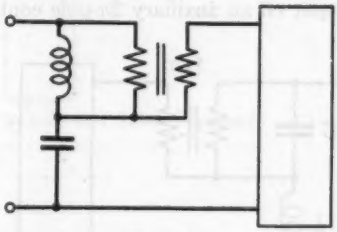


FIG. 4

FIG. 3. Reduction of a two-pole according to Brune
FIG. 4. Equivalent circuit to fig. 3

(c) *The minimum lies at finite frequency σ.* In this case the impedance matrix' after splitting off R, generally has no zero at $\lambda = j\sigma$. To be able to extend Brune's two-pole method in this case, let us recapitulate it.

Brune's method. At $\lambda = j\sigma$ the impedance function, after splitting off R, is reactive, $Z = jX$. If X is negative, Brune puts $X = \sigma L_1$, where L_1 is a negative inductance. To develop a zero of the impedance function at $\lambda = j\sigma$, Brune splits off L_1 in series. The impedance function thereby remains positive and a pole at infinity is introduced which raises the order of the two-pole by one. Then the zero is split off by a series connection of an inductance and a capacitance, in parallel to the two-pole. This lowers the order by two. Finally the pole at infinity is split off by a series inductance, lowering the order by one (fig. 3). The three steps together lower the order by two. Between the inductances there is a relation which makes it possible to replace the circuit of fig. 3 by that of fig. 4, containing one (positive) inductance, a capacitance and an ideal transformer.

If the reactance X is positive Brune puts $X = -1/\sigma C_1$, where C_1 is a negative capacitance. He now develops a zero of the admittance function by splitting off C_1 in parallel. We can, however, also in this case proceed by developing a zero of the impedance function, by splitting off C_1 in series.[18] Then, by a pro-

[18] W. Cauer, reference 3, p. 434.

MATHEMATICS GRADUATE LIBRARY

cedure quite analogous to that described above, we arrive at the circuit of fig. 5. By replacing this circuit by the equivalent circuit of fig. 6 and combining the second transformer with the reduced two-pole we see that also for positive X the circuit of fig. 4 can be used to reduce the two-pole. For positive X it is this procedure that we shall extend to the $2n$-pole.

Extension of Brune's method. Generally, for $n > 1$ it is not possible to develop a zero of the impedance matrix of a $2n$-pole at $\lambda = j\sigma$ by splitting off a reactance in series with one of the terminal pairs. Therefore Gewertz,[19] in the case $n = 2$, splits off reactances in series with each of the terminal pairs. The order of the four-pole is thereby raised by two, contrary to Brune's method, so this is not an appropriate way to extend it. This is confirmed by the fact that Gewertz does not proceed along this way but develops a different method, which cannot be regarded as an extension of Brune's method. To develop a zero of the impedance matrix at $\lambda = j\sigma$ whilst raising the order of the $2n$-pole by one, it is necessary to split off an auxiliary $2n$-pole containing one reactance. This can be achieved

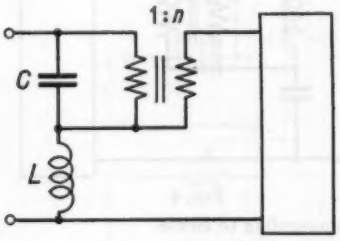


FIG. 5

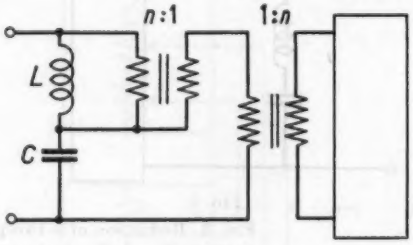


FIG. 6

FIG. 5. Reduction of a two-pole by a circuit similar to fig. 4
FIG. 6. Equivalent circuit to fig. 5

by splitting off a series $2n$-pole whose impedance matrix has rank one, consisting of a reactance X and $n - 1$ ideal transformers (section 7). Fig. 7 illustrates this for the case $n = 3$.

Just as Brune does, we must distinguish between negative and positive values of X. If X is negative we put $X = \sigma L_1$, where L_1 is a negative inductance. Then by splitting off the series $2n$-pole the impedance matrix remains positive. The matrix now has zeros at $\lambda = \pm j\sigma$ and a pole at $\lambda = \infty$. The zeros are generally of rank one (section 5) and can thus generally be split off by a parallel $2n$-pole consisting of an inductance and a capacitance in series and $n - 1$ ideal transformers. This lowers the order by two (section 6). Then the pole at infinity of the remaining impedance matrix is split off by a series $2n$-pole consisting of an inductance and $n - 1$ ideal transformers, lowering the order by one. The three steps together lower the order by two. For $n = 3$ they are illustrated by fig. 8. Between the inductances and the transformation ratios there are relations (section 8) which make it possible to replace the circuit of fig. 8 by

[19] Reference 5, Part X.

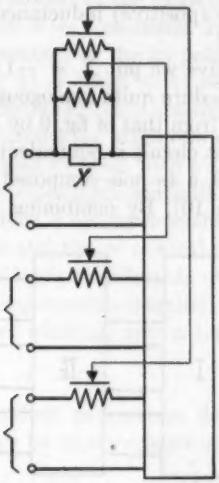


FIG. 7. Development of a zero of a $2n$-pole at a finite frequency

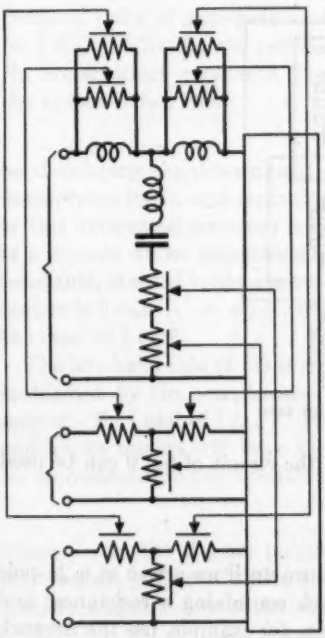


FIG. 8

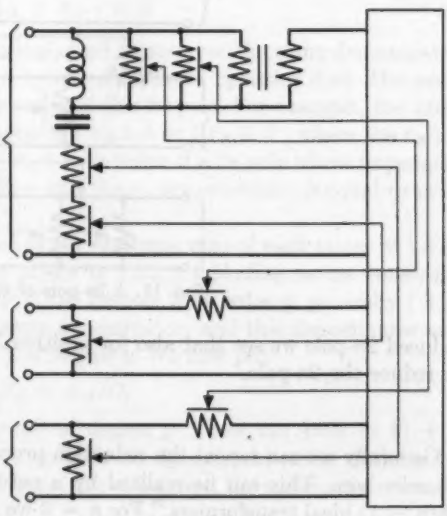


FIG. 9

FIG. 8. Reduction of a $2n$-pole; extension of fig. 3
FIG. 9. Equivalent circuit to fig. 8; extension of fig. 4

that of fig. 9, containing one (positive) inductance, a capacitance and $2n - 1$ ideal transformers (section 9).

If the reactance X is positive we put $X = -1/\sigma C_1$, where C_1 is a negative capacitance. Then, by a procedure quite analogous to that described above, we arrive at a circuit that arises from that of fig. 9 by interchanging the inductance and the capacitance. This last circuit is equivalent to that of fig. 10, consisting of a $4n$-pole like that in fig. 9, a $4n$-pole composed of n^2 ideal transformers, and the reduced $2n$-pole (section 10). By combining the second $4n$-pole with the

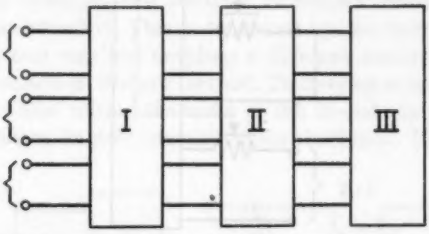


FIG. 10. Extension of fig. 6; I: a $4n$-pole like that in fig. 9, II: a $4n$-pole composed of n^2 ideal transformers, III: reduced $2n$-pole.

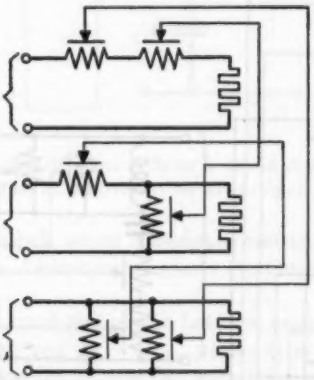


FIG. 11. A $2n$-pole of order zero

reduced $2n$-pole we see that also for positive X the circuit of fig. 9 can be used to reduce the $2n$-pole.

Generally we can repeat the reduction procedure until we arrive at a $2n$-pole of order zero. This can be realized by a network containing n resistances and $\frac{1}{2}n(n-1)$ ideal transformers.[20] For $n = 3$ we can, for example, use the network of fig. 11. We see that in this way the $2n$-pole is realized by a network containing

[20] W. Cauer, reference 3, p. 190.

the minimum number of reactances. In the cases (a) and (b), to lower the order of the $2n$-pole by one we need $n + 1$ elements. In case (c), to lower the order by two we need $2n + 2$ elements. For the $2n$-pole of order p we thus arrive at a network containing $p(n + 1) + \frac{1}{2}n(n + 1)$ elements. This is precisely the number of independent parameters defining a general $2n$-pole of order p (section 4). Thus the network also contains the minimum number of elements.

In special cases it may be that we start with a singular matrix or that, in the procedure described above, we come upon a singular matrix. As a singular matrix has no inverse, we cannot reduce it by the procedure. A singular matrix of rank r can be realized by a $2r$-pole and $r(n - r)$ ideal transformers.[21] A singular impedance (admittance) matrix leads to a $2n$-pole of which $n - r$ terminal pairs each have r transformer windings in series (parallel) and of which r terminal pairs each have $n - r$ transformer windings and a terminal pair of the $2r$-pole in parallel (series).

4. The number of independent parameters defining a general $2n$-pole of order p. Let a $2n$-pole be given by its impedance matrix

$$\| Z_{ik} \| \qquad (i, k = 1,2, \cdots, n).$$

To investigate the order of the $2n$-pole we connect its terminal pairs to the terminal pairs of a $2n$-pole containing no reactances, whose impedance matrix be $\| R_{ik} \|$. The Z_{ik} are rational functions of the frequency parameter λ; the R_{ik} are arbitrary constants. The differential equation of the free oscillations of the system arises from

$$| Z_{ik} + R_{ik} | = 0 \qquad (3)$$

by developing the determinant, bringing all terms over the same denominator, multiplying by it, and replacing λ by the differential operator d/dt. The order of this differential equation is the order of the $2n$-pole. For example, the order of a $2n$-pole whose impedance matrix is $\| r_{ik}\lambda \|$ or $\| r_{ik}/\lambda \|$, where the r_{ik} are constants, is equal to the rank of $\| r_{ik} \|$; the order of a $2n$-pole whose impedance matrix is $\| r_{ik}\lambda/(\lambda^2 + \sigma^2) \|$, where σ and the r_{ik} are constants, is equal to twice the rank of $\| r_{ik} \|$.

The left-hand side of (3) is equal to the algebraic sum of each minor of $| Z_{ik} |$ multiplied by the complementary minor of $| R_{ik} |$, including terms consisting only of $| Z_{ik} |$ and of $| R_{ik} |$. Thus, if the $2n$-pole is of order p, generally $| Z_{ik} |$ and all its minors will have the same denominator, and this denominator and all numerators are polynomials in λ of degree p. We now put

$$Z_{ik} = A_{ik}/C,$$

where C and the A_{ik} are polynomials of degree p. These are $\frac{1}{2}n(n + 1) + 1$ polynomials, each containing $p + 1$ coefficients. As only the ratios of the polynomials, and therefore of the coefficients, are of interest, the polynomials contain

$$(p + 1)\{\tfrac{1}{2}n(n + 1) + 1\} - 1 = p\{\tfrac{1}{2}n(n + 1) + 1\} + \tfrac{1}{2}n(n + 1) \qquad (4)$$

[21] W. Cauer, reference 3, p. 141. R. Leroy, reference 10, Théorème IV.

parameters. However, these parameters are not independent, for $| Z_{ik} |$ and all its minors must have C as denominator. Thus $| A_{ik} |$ must contain C^{n-1} as a factor; the minors of $| A_{ik} |$ of order $n-1$ must contain C^{n-2} as a factor; $\cdots$; the minors of $| A_{ik} |$ of order 2 must contain C as a factor.

We shall now show that to find the relations between the parameters it is generally only necessary to investigate the minors of $| A_{ik} |$ of order 2. Let B be a determinant of order s. The elements of the inverse determinant are formed by the minors of order $s-1$ divided by B. As the inverse determinant is equal to $1/B$, the determinant D formed by the afore-mentioned minors is equal to B^{s-1}. Let B contain r factors C, then D will generally contain $r(s-1)$ factors C. (It is only in the special case that C has multiple zeros that B^{s-1} can contain more than $r(s-1)$ factors C.) Supposing now that all minors of B of order $s-1$ contain C^{s-2} as a factor, then D will contain at least $s(s-2)$ factors C. Thus, generally, $r(s-1) \geqq s(s-2)$, or $r \geqq s-1-1/(s-1)$. As r and s are whole numbers this leads to

$$r \geqq s-1 \qquad \text{for } s \geqq 3. \tag{5}$$

If the minors of $| A_{ik} |$ of order 2 contain C as a factor, then by applying (5) for $s=3$ we find that the minors of $| A_{ik} |$ of order 3 generally contain C^2 as a factor. By applying (5) for $s=4$ we find next that the minors of $| A_{ik} |$ of order 4 generally contain C^3 as a factor, and so on. Therefore it is generally only necessary to investigate the condition that the minors of $| A_{ik} |$ of order 2 contain C as a factor.

We shall now show that it is only necessary to investigate $\frac{1}{2}n(n-1)$ of these minors, e.g. all minors of order 2 that have A_{11} in its upper left-hand corner. If we know that these minors contain C as a factor, then generally all minors of order 2 will contain C as a factor. To prove this we consider the following three minors:

$$\begin{vmatrix} A_{11} & A_{1j} \\ A_{i1} & A_{ij} \end{vmatrix}, \qquad \begin{vmatrix} A_{11} & A_{1k} \\ A_{i1} & A_{ik} \end{vmatrix}, \qquad \begin{vmatrix} A_{1j} & A_{1k} \\ A_{ij} & A_{ik} \end{vmatrix}.$$

If the first two minors contain C as a factor, at a zero of C we have

$$A_{ij}A_{11} - A_{1j}A_{i1} = 0,$$
$$A_{ik}A_{11} - A_{1k}A_{i1} = 0.$$

Therefore if A_{11} and A_{i1} are not both zero, which generally will not be the case, the third minor will be zero too. (If A_{11} and A_{i1} are both zero at a zero of C, the first two minors are zero, and we must separately demand the third minor to be zero too. This means that three polynomials must have the zero of C, instead of only two in the general case, so this leads to a less general $2n$-pole.) Thus, if the first two minors contain C as a factor, the third will generally also contain C as a factor. By extending this reasoning we come to the above-mentioned property.

To find the conditions that must exist between the coefficients of the poly-

nomials C and A_{ik} of degree p in order that these may define a $2n$-pole of order p, we have generally only to express that the afore-mentioned $\frac{1}{2}n(n-1)$ minors of order 2 contain C as a factor. This means that every zero of C is also a zero of these minors. As C is of degree p, this leads to p conditions for every minor, and so in total to $p\{\frac{1}{2}n(n-1)\}$ conditions. The number of independent parameters defining a general $2n$-pole of order p is found by subtracting from (4) the number of these conditions, and so it amounts to $p(n+1)+\frac{1}{2}n(n+1)$.

5. Poles of rank one. Poles of an impedance (or admittance) matrix of a $2n$-pole of a given order are generally of rank one. Let us consider a pole at $\lambda = 0$. Then the polynomial C of section 4 will have a zero at $\lambda = 0$ which generally will be simple. As seen in section 4, every minor of $|\, A_{ik}\,|$ of order 2 contains C as a factor and therefore will contain λ as a factor. This means that the determinants formed by the constant terms of the polynomials constituting the aforementioned minors are zero. These constant terms are proportional to the residues of the pole, and so the pole is generally of rank one. This is also true for poles at other values of λ.

6. The lowering of the order. We prove the following

THEOREM. *If two $2n$-poles whose impedance matrices have no poles in common are connected in series, the order of the combined $2n$-pole is equal to the sum of the orders of the component $2n$-poles.*

Let the $2n$-poles have impedance matrices $\|\, Z'_{ik}\,\|$ and $\|\, Z''_{ik}\,\|$, let their orders be p' and p'', respectively, and let the common denominators of $|\, Z'_{ik}\,|$ and its minors and of $|\, Z''_{ik}\,|$ and its minors be C' and C'', respectively. Then $|\, Z'_{ik}\,|$ and its minors, zero order included, multiplied by C', and $|\, Z''_{ik}\,|$ and its minors, zero order included, multiplied by C'' are polynomials in λ of highest degree p' and p'', respectively. The impedance matrix of the combined $2n$-pole is $\|\, Z'_{ik}+Z''_{ik}\,\|$. Since $\|\, Z'_{ik}\,\|$ and $\|\, Z''_{ik}\,\|$ have no poles in common, C' and C'' have no common factors. Thus the common denominator of $|\, Z'_{ik}+Z''_{ik}\,|$ and its minors is $C'C''$. The order of the combined $2n$-pole is therefore equal to the highest degree of $|\, Z'_{ik}+Z''_{ik}\,|$ and its minors, zero order included, multiplied by $C'C''$. Now $|\, Z'_{ik}+Z''_{ik}\,|$ and its minors can be developed into an algebraic sum of products of $|\, Z'_{ik}\,|$, $|\, Z''_{ik}\,|$, and their minors. From this it follows that the order of the combined $2n$-pole cannot be higher than $p'+p''$.

The order of the combined $2n$-pole cannot be lower than $p'+p''$ either, which can be shown as follows. Since $\|\, Z'_{ik}\,\|$ and $\|\, Z''_{ik}\,\|$ do not both have poles at $\lambda = \infty$, let us suppose that the first has no such pole. Then C' will be of degree p'. As the second component $2n$-pole is of order p'', $|\, Z''_{ik}\,|$ or some minor, zero order included, multiplied by C'' will be a polynomial of degree p''. Between the minors, $|\, Z''_{ik}\,|$ included, having this property we choose one having the lowest order. For example, if C'' is of degree p'' we choose the minor of zero order 1; if C'' is of degree $p''-1$ we choose some minor of first order, Z''_{ik}, whose numerator is of order p''; and so on. We now consider a minor of $|\, Z'_{ik}+Z''_{ik}\,|$, nth and zero order included, whose order is equal to the afore-mentioned lowest

order and that, after developing into an algebraic sum, contains the aforementioned minor of $| Z''_{ik} |$ as a term. All other terms of this development contain a minor of $| Z''_{ik} |$ of lower order, whose numerator is therefore of a degree lower than p''. Thus the considered minor of $| Z'_{ik} + Z''_{ik} |$ multiplied by $C'C''$ contains, after development, one term of degree $p' + p''$, which proves our statement.

As mentioned in section 4, the order of a $2n$-pole whose impedance matrix is $\| r_{ik}/\lambda \|$ is equal to the rank of $\| r_{ik} \|$. From this and our theorem it follows that if the impedance matrix of a $2n$-pole has a pole at $\lambda = 0$ of rank one, after splitting off this pole the order of the $2n$-pole has been lowered by one, the same applying for the splitting off of a pole of the admittance matrix and for a pole at $\lambda = \infty$. If there are poles at $\lambda = \pm j\sigma$ of rank one, by splitting them off the order of the $2n$-pole is lowered by two.

If the poles considered are of a rank higher than one, by splitting them off the order of the $2n$-pole is lowered by more than one or two. In the case where, in the extension of Brune's method (section 3), poles arise in the admittance matrix at $\lambda = \pm j\sigma$ of a rank higher than one, we do not split off these poles completely but confine ourselves to splitting off poles of rank one, thereby lowering the order of the $2n$-pole by two. That this procedure is always possible follows from a lemma by Oono[22] on symmetrical matrices the elements of which are real numbers. Such a matrix is said to be positive if the associated quadratic form is positive, definite or semi-definite. Oono showed that a positive matrix of rank r can be expressed as the sum of r positive matrices of rank 1. By condition (d) for a positive real matrix (section 2), the residues of a pole on the imaginary axis of a positive real matrix constitute the elements of such a positive matrix. Therefore, if a matrix has a pole at $\lambda = \pm j\sigma$ of rank r, we can first split off this pole completely by means of a matrix of rank r, thereby lowering the order of the $2n$-pole by $2r$. Then we can, by Oono's lemma, split the matrix of rank r into the sum of r matrices of rank one. Then we add $r - 1$ of these matrices together, which results in a matrix of rank $r - 1$. By adding this matrix to that of the reduced $2n$-pole, the order of the $2n$-pole is raised by $2r - 2$ and thus becomes equal to the order of the original $2n$-pole diminished by two.

7. The development of a zero of a matrix at $\lambda = j\sigma$. In this section we consider symmetrical matrices the elements of which are complex numbers. Such a matrix we say to be positive if the quadratic form associated with the real parts of its elements is positive, definite or semi-definite. The impedance matrix of a passive $2n$-pole is positive in this sense at each point of the imaginary axis. To develop a zero of the impedance matrix $\| Z_{ik} \| = \| R_{ik} + jX_{ik} \|$ on the imaginary axis we began in section 3 by decreasing R_{11} until $| R_{ik} | = 0$ at some point of the imaginary axis. We now suppose this point to be $\lambda = j\sigma$. The impedance matrix of the $2n$-pole consisting of one reactance and $n - 1$ ideal transformers, as illustrated by fig. 7 for the case $n = 3$, is of rank one and has imaginary elements. The way to develop a zero of a positive matrix for which $| R_{ik} | = 0$ by subtracting a matrix of rank one with imaginary elements, is shown by the following

THEOREM.[23] *If a positive matrix $\| Z_{ik} \|$ of order n has determinant zero, there exists a set of real numbers β_k not all zero, such that*

[22] Reference 9, p. 22.

[23] W. Peremans, H. J. A. Duparc, C. G. Lekkerkerker, Proc. Kon. Ned. Akad. Wetensch.

$$\sum_{k=1}^{n} Z_{ik}\beta_k = 0 \qquad \text{for } i = 1, \cdots, n.$$

A symmetrical matrix of rank one with imaginary elements can be written as

$$\| j\alpha h_i h_k \| \; ; \; h_i \, , \; h_k \text{ real}; \; \alpha = \pm 1.$$

The h's and α must now be so determined that we can write

$$\sum_{k=1}^{n} (R_{ik} + jX_{ik} - j\alpha h_i h_k)\beta_k = 0. \tag{6}$$

The real parts of (6) lead to

$$\sum_{k=1}^{n} R_{ik}\beta_k = 0.$$

Since $| R_{ik} | = 0$, there exist real β's that are not all zero and that satisfy these equations. The imaginary parts of (6) lead to

$$\alpha h_i \sum_{k=1}^{n} h_k \beta_k = \sum_{k=1}^{n} X_{ik}\beta_k. \tag{7}$$

By multiplying by β_i and summing over i we get

$$\alpha \left(\sum_{k=1}^{n} h_k \beta_k \right)^2 = \sum_{i,k=1}^{n} X_{ik}\beta_i\beta_k.$$

This determines the sign of α and the value of $\sum_{k=1}^{n} h_k\beta_k$. From (7) we can then find the values of the h's.[24] From (6) it follows that $| R_{ik} + jX_{ik} - j\alpha h_i h_k | = 0$, so that we have developed the required zero of the impedance matrix.

8. The relations between the inductances and the transformation ratios in fig. 8. The procedure illustrated by fig. 8, part of which, for an arbitrary value of n, is redrawn in fig. 12, starts from a $2n$-pole whose impedance matrix has no pole at $\lambda = \infty$. Then such a pole is introduced by subtracting a matrix $\| p_i p_k \lambda L_1 \|$ (with $p_1 = 1$). In this way zeros are developed at $\lambda = \pm j\sigma$. These are removed by subtracting from the admittance matrix a matrix $\| n_i n_k / [\lambda L_2 + (1/\lambda C)] \|$ (with $n_1 = -1$),[25] where $\sigma^2 L_2 C = 1$. We are interested in the pole at $\lambda = \infty$ of the ensuing impedance matrix.

If $\| Z_{ik} \|$ is an impedance matrix and $\| Y_{ik} \|$ the corresponding admittance matrix, we have

$$Z_{ik} = y_{ik}/| Y_{ik} |, \qquad Y_{ik} = z_{ik}/| Z_{ik} |,$$

Amsterdam, A **55**, 24, 1952. Leroy, reference 10, Théorème II, has proved the theorem for the special case of a positive real matrix, in the sense of section 2, whose determinant is zero for all values of λ.

[24] We suppose the h's to be finite. The case where the h's are infinite can be considered as a limiting case (see note 26).

[25] The minus sign is arbitrary. It somewhat simplifies the equations of the next section.

where y_{ik} is the cofactor of Y_{ik} in $|\,Y_{ik}\,|$, and z_{ik} the cofactor of Z_{ik} in $|\,Z_{ik}\,|$. Further we know that $|\,Z_{ik}\,| \cdot |\,Y_{ik}\,| = 1$. We now put

$$Z'_{ik} = Z_{ik} - p_i p_k \lambda L_1 = y'_{ik}/|\,Y'_{ik}\,|.$$

Then

$$|\,Y'_{ik}\,| = \frac{1}{|\,Z'_{ik}\,|} = \frac{1}{|\,Z_{ik}\,| - \sum_{i,k=1}^{n} p_i p_k \lambda L_1 z_{ik}},$$

$$y'_{ik} = (Z_{ik} - p_i p_k \lambda L_1)\,|\,Y'_{ik}\,|.$$

For $\lambda \to \infty$, putting

$$\left(\sum_{i,k=1}^{n} p_i p_k L_1 z_{ik}\right)_{\infty} = Q,$$

we get

$$|\,Y'_{ik}\,|_{\infty} = -1/\lambda Q, \qquad y'_{ik_{\infty}} = p_i p_k L_1/Q.$$

We further put

$$Y''_{ik} = Y'_{ik} - n_i n_k/[\lambda L_2 + (1/\lambda C)].$$

Since for $\lambda \to \infty$ the last term tends to zero, we have

$$y''_{ik_{\infty}} = y'_{ik_{\infty}} = p_i p_k L_1/Q.$$

Further

$$|\,Y''_{ik}\,| = |\,Y'_{ik}\,| - \sum_{i,k=1}^{n} \frac{n_i n_k}{\lambda L_2 + (1/\lambda C)} \cdot y'_{ik}.$$

For $\lambda \to \infty$ this results in

$$|\,Y''_{ik}\,|_{\infty} = -\frac{1}{\lambda Q} - \frac{L_1}{\lambda Q L_2} \sum_{i,k=1}^{n} p_i p_k n_i n_k = -\frac{L_2 + L_1 \left(\sum_{k=1}^{n} p_k n_k\right)^2}{\lambda Q L_2}.$$

For the corresponding impedance matrix we get

$$Z''_{ik\infty} = \frac{y''_{ik\infty}}{|\,Y''_{ik}\,|_{\infty}} = -\frac{p_i p_k \lambda L_1 L_2}{L_2 + \left(\sum_{k=1}^{n} p_k n_k\right)^2 L_1}.$$

This shows that the matrix has a pole at $\lambda = \infty$ of rank one. The pole can be removed by subtracting a matrix $\| p_i p_k \lambda L_3 \|$, where L_3 is given by

$$L_1 L_2 + L_2 L_3 + \left(\sum_{k=1}^{n} p_k n_k\right)^2 L_1 L_3 = 0. \tag{8}$$

Now $\sum_{k=1}^{n} p_k n_k = \sum_{k=2}^{n} p_k n_k - 1$. Thus if we put $\sum_{k=2}^{n} p_k n_k = A$, we can write

$$L_1 L_2 + L_2 L_3 + (A - 1)^2 L_1 L_3 = 0. \tag{8'}$$

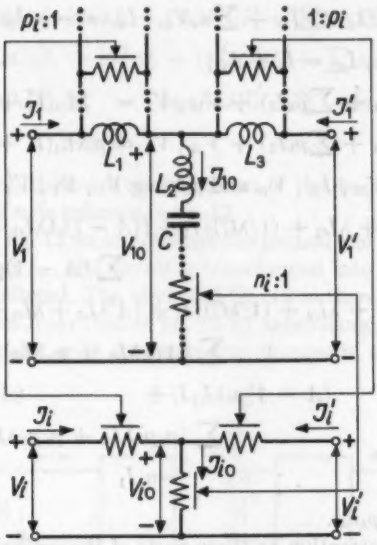


FIG. 12. The $4n$-pole from fig. 8

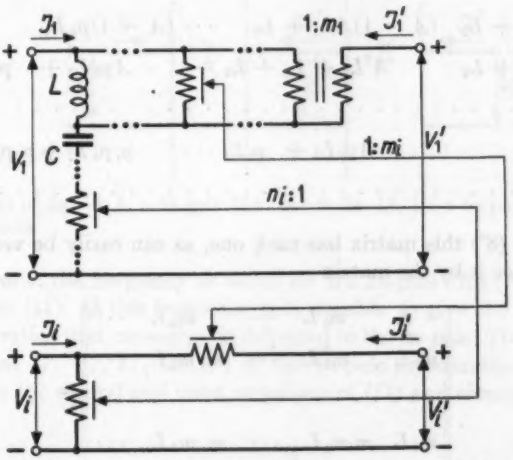


FIG. 13. The $4n$-pole from fig. 9; equivalent circuit to fig. 12.

9. The circuits of fig. 8 and fig. 9 are equivalent. *In this and the next section* $i, k = 2, \cdots, n$ *and* $\sum = \sum_{k=2}^{n}$.

The $4n$-pole in fig. 8 is, for an arbitrary value of n, redrawn in fig. 12. The equations of the $4n$-pole are

$$V_{10} = [\lambda L_2 + (1/\lambda C)]I_{10} + \sum n_k V_{k0}\,,\ I_{i0} = -\,n_i I_{10}\,;$$
$$I_{10} = I_1 + I_1'\,,\ I_{i0} = I_i + I_i'\,;$$
$$V_1 = \lambda L_1(I_1 + \sum p_k I_k) + V_{10}\,,\ V_1' = \lambda L_2(I_1' + \sum p_k I_k') + V_{10}\,,$$
$$V_i = p_i\lambda L_1(I_1 + \sum p_k I_k) + V_{i0}\,,\ V_i' = p_i\lambda L_2(I_1' + \sum p_k I_k') + V_{i0}\,.$$

By eliminating I_{10}, V_{10}, I_{i0}, V_{i0} and solving V_1, V_1', V_i', I_i, we find

$$V_1 = \{(A-1)^2\lambda L_1 + \lambda L_2 + (1/\lambda C)\}I_1 + \{(A-1)A\lambda L_1 + \lambda L_2 + (1/\lambda C)\}I_1' + \sum (A-1)p_k\lambda L_1\, I_k' + \sum n_k V_k,$$
$$V_1' = \{(A-1)A\lambda L_1 + \lambda L_2 + (1/\lambda C)\}I_1 + \{A^2\lambda L_1 + \lambda L_2 + \lambda L_2 + (1/\lambda C)\}I_1' + \sum (Ap_k\lambda L_1 + p_k\lambda L_2)\, I_k' + \sum n_k V_k, \quad (9)$$
$$V_i' = (A-1)p_i\lambda L_1 I_1 + (Ap_i\lambda L_1 + p_i\lambda L_2)\, I_1' + \sum (p_i p_k\lambda L_1 + p_i p_k\lambda L_2)I_k' + V_i,$$
$$I_i = -n_i I_1 - n_i I_1' - I_i',$$

where again $A = \sum p_k n_k$.

We now turn our attention to those parts of the coefficients of the equations that are proportional to λ. These form a matrix

$$\begin{Vmatrix} (A-1)^2 L_1 + L_2 & (A-1)AL_1 + L_2 & \cdots & (A-1)p_k L_1 & \cdots \\ (A-1)A L_1 + L_2 & A^2 L_1 + L_2 + L_2 & \cdots & Ap_k L_1 + p_k L_2 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ (A-1)p_i L_1 & Ap_i L_1 + p_i L_2 & \cdots & p_i p_k L_1 + p_i p_k L_2 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{Vmatrix}. \quad (10)$$

On account of (8′) this matrix has rank one, as can easily be verified. We can therefore replace it by the matrix

$$\begin{Vmatrix} L & m_1 L & \cdots & m_k L & \cdots \\ m_1 L & m_1^2 L & \cdots & m_1 m_k L & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ m_i L & m_i m_1 L & \cdots & m_i m_k L & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{Vmatrix},$$

where $L = (A-1)^2 L_1 + L_2 = -L_1 L_2/L_3$ is positive. The equations (9) thereby change into[26]

[26] The case mentioned in note 24 also leads to the equations (11). If L_1 becomes infinite, the p's, the n's, L_2, $(A-1)L_1$, and $L_1 + L_2$ appear to remain finite. Thus all elements of (10) remain finite and so also L and the m's.

$$
\begin{aligned}
V_1 &= [\lambda L + (1/\lambda C)]I_1 + [m_1\lambda L + (1/\lambda C)]I_1' + \sum m_k\lambda L I_k' + \sum n_k V_k,\\
V_1' &= [m_1\lambda L + (1/\lambda C)]I_1 + [m_1^2\lambda L + (1/\lambda C)]I_1' + \sum m_1 m_k\lambda L I_k' + \sum n_k V_k,\\
V_i' &= \quad m_i\lambda L I_1 + \quad m_i m_1\lambda L I_1' + \sum m_i m_k\lambda L I_k' + \quad V_i, \qquad (11)\\
I_i &= \quad -n_i I_1 \quad\quad\quad -n_i I_1' \quad\quad\quad -I_i' \quad .
\end{aligned}
$$

As can easily be seen, these are the equations of the $4n$-pole in fig. 9, which, for an arbitrary value of n, is redrawn in fig. 13.

If in the circuit of fig. 12 we interchange the primed and the unprimed terminal pairs and also L_1 and L_3, the circuit is transformed into itself. The relation (δ) thereby remains unaltered. The circuit of fig. 12 is therefore also equivalent to the circuit that arises from that of fig. 13 by interchanging the primed and the unprimed terminal pairs, with corresponding changes in the values of L and the m's.

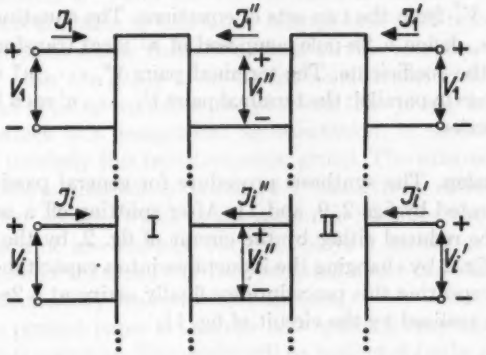


FIG. 14. Part of fig. 10; I: a $4n$-pole like that in fig. 13, II: a $4n$-pole composed of n^2 ideal transformers.

The value of σ, the frequency at which for the $2n$-pole $|R_{ik}| = 0$, can be determined from (11). At this frequency it is possible to give the currents to the $2n$-pole such ratios that no energy is delivered to the $2n$-pole. This means that it is possible that I_1', V_1', I_i', and V_i' of the $4n$-pole are simultaneously zero. If we put this in the second and third equations of (11) and eliminate I_1, V_i, we find

$$\sigma^2 = \frac{1}{(m_1 - \sum m_k n_k)LC}. \qquad (12)$$

This shows $m_1 - \sum m_k n_k$ to be positive.

10. The circuit arising from that of fig. 9 by interchanging inductance and capacitance is equivalent to the circuit of fig. 10. We consider the $4n$-pole whose equations arise from (11) by interchanging λL and $1/\lambda C$. This $4n$-pole is equiva-

lent to the $4n$-pole of fig. 14 if I is defined by equations that arise from (11) by replacing

$$I_1', V_1', I_i', V_i' \text{ by } I_1'', V_1'', I_i'', V_i'';$$

$$m_1 \text{ by } \frac{1 - \sum m_k n_k}{m_1 - \sum m_k n_k}, \qquad m_i \text{ by } -\frac{m_i}{m_1 - \sum m_k n_k};$$

and if II is defined by the equations

$$\begin{aligned} I_1'' &= m_1 I_1' + \sum m_k I_k', \\ I_i'' &= n_i(1 - m_1) I_1' + \sum (\delta_{ik} - n_i m_k) I_k', \\ V_1' &= m_1 V_1'' + \sum n_k(1 - m_1) V_k'', \\ V_i' &= m_i V_1'' + \sum (\delta_{ik} - m_i n_k) V_k'', \end{aligned} \tag{13}$$

where $\delta_{ik} = 1$ for $i = k$, $\delta_{ik} = 0$ for $i \neq k$. This can be verified by eliminating I_1'', V_1'', I_i'', V_i'' from the two sets of equations. The equations (13), containing n^2 coefficients, define a $2n$-pole composed of n^2 ideal transformers whose ratios are equal to the coefficients. The terminal pairs $1'', \cdots, n''$ each have n transformer windings in parallel; the terminal pairs $1', \cdots, n'$ each have n transformer windings in series.

11. Conclusion. The synthesis procedure for general passive $2n$-poles is, for $n = 3$, illustrated by figs 2, 9, and 11. After splitting off a series resistance the $2n$-pole can be reduced either by the circuit of fig. 2, by the circuit that arises from that of fig. 2 by changing the inductance into a capacitance, or by the circuit of fig. 9. By repeating this procedure we finally arrive at a $2n$-pole of zero order which can be realized by the circuit of fig. 11.

PHILIPS RESEARCH LABORATORIES
EINDHOVEN-NETHERLANDS

(Received May 23, 1952)

SPINOR THEORY OF FOUR-TERMINAL NETWORKS

W. T. Payne

1. Introduction. In the theory of linear four-terminal networks, it is sometimes considered desirable to have a geometrical scheme that will give a picture of network operation and will help in solving network problems. All the geometrical schemes that have been devised hitherto[1] appear to be based on the linear fractional transformations of a single complex variable, which is taken to represent the impedance, or ratio of complex voltage to complex current, at the input or output of the network. Such a scheme has two shortcomings: first, it fails to give a completely detailed account of network operation, since it describes only the transformation of the ratio of voltage to current rather than the transformation of voltage and current themselves; and second, it can be applied only to networks whose determinant is 1, and not, for example, to amplifier circuits. For these reasons, it seems desirable to find a new scheme that shall be based on the group of all homogeneous linear transformations of two complex variables, a group that is larger than the group of linear fractional transformations of one complex variable and is better suited to the problem. The proper scheme is the theory of spinors, which is a geometrical representation, in three-dimensional Euclidean space, of precisely this transformation group. The subgroup of homogeneous linear transformations of determinant 1 (sometimes called the group C_2[2] or the unimodular group), which is the algebraic basis for the linear fractional transformations of one complex variable, plays a special part in spinor theory, just as it does in the theory of four-terminal networks, where it is the basis for networks that satisfy the reciprocity theorem.

The object of the present paper is to show how spinor theory can be applied to linear four-terminal networks. The study will be restricted to the steady state, although the spinor scheme appears to be broad enough so that this restriction is perhaps not really necessary. For the sake of mathematical completeness, negative resistances will not be excluded from consideration, although, of course, the reasoning does not depend in any essential way on their inclusion.

2. Elements of spinor theory. The basic part of spinor theory has recently been presented by the author in an elementary way,[3] and will now be briefly outlined and extended to meet the needs of the present application. A spinor can be described by saying that it is a geometric object in 3-space having a magnitude

[1] H. König, *Über die Abhängigkeit des Scheinwiderstandes eines symmetrischen Vierpols von der Belastung*, *Helvetica Physica Acta* **4**, 281 (1931). A Weissfloch, *Kreisgeometrische Vierpoltheorie*, *Hochfrequenztechnik und Elektroakustik* **61**, 100 (1943). J. van Slooten *Meetkundige beschouwingen in verband met de theorie der electrische vierpolen*, thesis, Delft, 1946.

[2] B. L. van der Waerden, *Die gruppentheoretische Methode in der Quantenmechanik*, Springer, Berlin, 1932. L. Brillouin, *Wave Propagation in Periodic Structures*, McGraw-Hill, New York, 1946.

[3] W. T. Payne, *Elementary Spinor Theory*, *American Journal of Physics* **20**, 253 (1952).

(S) and three Eulerian angles, as shown in Figure 1. The analytical theory of spinors is based on two complex components S_1 and S_2, defined by the equations

$$S_1 = (2S)^{\frac{1}{2}} \cos \tfrac{1}{2}\theta e^{i(\psi+\phi)/2}$$
$$S_2 = (2S)^{\frac{1}{2}} \sin \tfrac{1}{2}\theta e^{i(\psi-\phi)/2}. \tag{1}$$

If a spinor has specifications S, θ, ϕ, and ψ, then the vector in 3-space whose magnitude, polar angle, and azimuthal angle are equal to S, θ, and ϕ, respectively, may be called the associated vector[4] **P**. Its components and magnitude are related to the spinor's components as follows:

$$P_1 = \tfrac{1}{2}(S_1S_2^* + S_2S_1^*), \qquad P_2 = -\tfrac{1}{2}i(S_1S_2^* - S_2S_1^*),$$
$$P_3 = \tfrac{1}{2}(S_1S_1^* - S_2S_2^*), \qquad P = \tfrac{1}{2}(S_1S_1^* + S_2S_2^*). \tag{2}$$

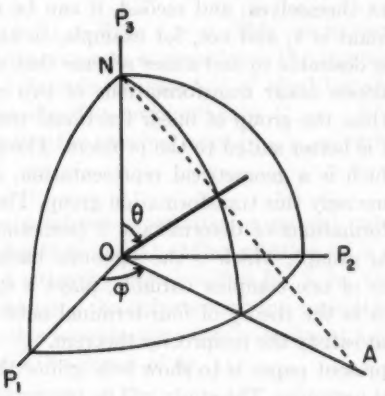


FIG. 1. Spinor

The direction of a spinor can be specified by the ratio of its components, S_1/S_2, which, according to Eqs. (1), is equal to $\cot \frac{1}{2}\theta e^{i\phi}$ and thus depends only on θ and ϕ, but not on ψ. We shall call this ratio σ. Evidently the quantity $\cot \frac{1}{2}\theta e^{i\phi}$ may be used to specify the direction of a ray, as well as that of a spinor. Two rays whose σ's are the negatives of each other have their angle bisected by the P_3-axis, and conversely.

A geometric picture of the relation between a spinor (or ray) and the associated σ-value can be obtained as follows. Describe a sphere of unit radius about the origin; project each point of the sphere onto the P_1P_2-plane by stereographic projection from the north pole of the sphere; and transfer the value of σ from each point of the sphere to its projection on the plane. Then elementary geometry shows that the different values of σ will be laid out on the P_1P_2-plane according to the usual rules for plotting complex numbers in the complex plane.

[4] See Reference 3, Eq. (17). Note change in notation. The transverse associated vectors are not used in the present paper.

The stereographic projection is shown in Figure 2, in which N is the north pole and A is the projected point, and OA is equal to cot $\frac{1}{2}\theta$, if the sphere has unit radius.

On looking back over this section, we see that a spinor has four real specifications, the associated vector three (the angle ψ having been eliminated), and the direction ratio σ only two. In the application of spinor theory to four-terminal networks, as will be shown later, the spinor components represent complex current and complex voltage, the associated vector represents power, and σ represents impedance.

3. The homogeneous linear transformations of a spinor. It will be shown at the end of this section that all non-singular homogeneous linear transformations of the components of a spinor can be made up multiplicatively out of three standard types of transformations, which are rotations, unimodular (i.e. having determinant equal to unity) Hermitian transformations, and numerical multiples of the identity transformation. These three types will now be studied briefly.

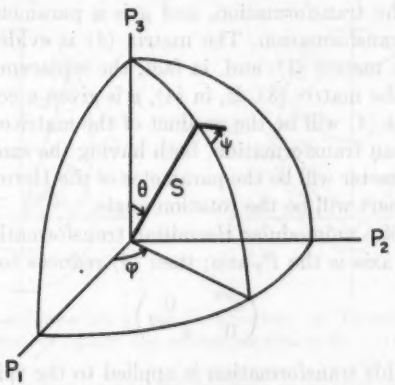


FIG. 2. Stereographic projection

Rotations have already been discussed in the previous paper.[3] If a spinor is rotated through the angle κ about an axis whose direction angles are α, β, γ, then its components undergo a homogeneous linear transformation whose matrix is

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \cos \tfrac{1}{2}\kappa + \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix} \sin \tfrac{1}{2}\kappa \cos \alpha + \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \sin \tfrac{1}{2}\kappa \cos \beta + \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix} \sin \tfrac{1}{2}\kappa \cos \gamma. \tag{3}$$

This matrix is a unimodular unitary[5] matrix. It is not difficult to prove that every two-rowed unimodular unitary matrix can be written in the form (3), with

[5] For definitions, see any book on matrices or on modern algebra.

real values for κ, α, β, γ. In proving this, it is helpful to note that a two-rowed unimodular unitary matrix has the sum of its diagonal elements real and between -2 and 2, and their product real and between 0 and 1.

As for unimodular Hermitian transformations, a unimodular Hermitian[5] matrix can always be expressed in the form

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \cosh \tfrac{1}{2}\mu + \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \sinh \tfrac{1}{2}\mu \cos \alpha + \begin{pmatrix} 0 & i \\ -i & 0 \end{pmatrix} \sinh \tfrac{1}{2}\mu \cos \beta + \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \sinh \tfrac{1}{2}\mu \cos \gamma, \tag{4}$$

or the negative of this, with real values for μ, α, β, γ. This statement can be proved by using the facts that a unimodular Hermitian matrix has the sum of its diagonal elements real and not lying between -2 and 2, and their product real and not less than 1. The direction specified by the angles α, β, γ may be called the axis of the transformation, and μ is a parameter that specifies the magnitude of the transformation. The matrix (4) is evidently very similar in form to the rotation matrix (3); and, in fact, the replacement in (4) of μ by $i\kappa$ will yield precisely the matrix (3). If, in (4), μ is given a complex value, then it is easily verified that (4) will be the product of the matrices of a rotation and a unimodular Hermitian transformation, both having the same axis; the real part of the complex parameter will be the parameter of the Hermitian transformation and the imaginary part will be the rotation angle.

To get a picture of a unimodular Hermitian transformation of a spinor, let us consider one whose axis is the P_3-axis; then (4) reduces to

$$\begin{pmatrix} e^{\mu/2} & 0 \\ 0 & e^{-\mu/2} \end{pmatrix}, \tag{5}$$

with μ real. When this transformation is applied to the spinor components (1), it evidently leaves the values of ϕ and ψ unchanged, and changes only S and θ. We can see how S and θ are changed by investigating the effect of the transformation (5) on the associated vector. The use of Eqs. (2) yields the following transformation equations for this vector, the primed symbols representing the values after transformation:

$$\begin{aligned} P_1' &= P_1, \qquad P_2' = P_2 \\ P_3' &= P_3 \cosh \mu + P \sinh \mu \\ P' &= P_3 \sinh \mu + P \cosh \mu. \end{aligned} \tag{6}$$

It may be noted that (6) is a Lorentz transformation, though this fact is of no particular significance here. The first two of Eqs. (6) show that the transformation (5) causes the blade[6] of a spinor to move along a line parallel to the P_3-axis One of the equations inverse to (6) will be

$$P = P' \cosh \mu - P_3' \sinh \mu, \tag{7}$$

[6] See Reference 3, Section 2.

and if P is held constant this is the equation of a prolate spheroid with one focus at the origin, major axis along the P_3-axis, and with semimajor axis a equal to $P \cosh \mu$, semiminor axis b equal to P, and focal distance c equal to $P \sinh \mu$. (To show this, take the standard form of the prolate spheroid equation, which is $P'a(1 - \epsilon \cos \theta') = b^2$; note that this can be written $P'a - P_3'c = b^2$, since $c = \epsilon a$ and $P_3' = P' \cos \theta'$; and compare with (7).) We conclude that all spinors initially terminating on a sphere will, after the transformation (5) is applied to them, terminate on the prolate spheroid just described. Figure 3a shows a picture of this; the dotted line is the path along which the spinor blade moves, $\mathbf{S}'$ being the spinor into which $\mathbf{S}$ is transformed.

The transformation (5) transforms the P_1P_2-plane into a cone whose vertex is at the origin and which intersects each of the prolate spheroids just mentioned around its middle. Every plane parallel to the P_1P_2-plane is transformed by (5) into one sheet of a two-sheet hyperboloid of rotation having a focus at the origin.

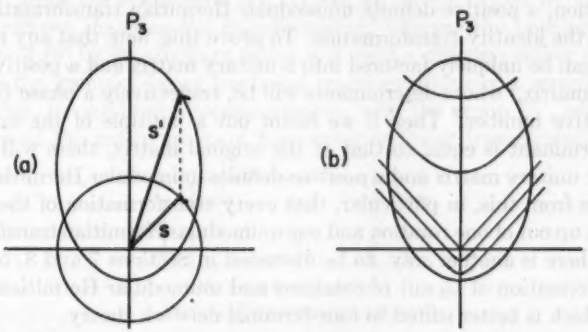


FIG. 3. Hermitian transformation in the P_3-direction. (a) Transformation of a spinor. (b) Transformation of the P_1P_2-plane and planes parallel to it.

These statements are readily proved by analytic geometry, if one uses the equation inverse to the third of Eqs. (6). The arrangement of cone and hyperboloids is shown in Figure 3b. Since the focal distance of an ellipsoid of semiminor axis P is equal to $P \sinh \mu$, it follows that the half-angle of the cone has its cotangent equal to $\sinh \mu$, so that the cotangent of half the half-angle must be equal to e^{μ}.

This geometrical description of the transformation (5) can be applied to any other positive-definite[5] Hermitian transformation of determinant 1, by merely reorienting the common axis of the prolate spheroid and the cone and hyperboloids so that it will coincide with the axis of the given transformation. The analytical justification for this statement consists in the fact that the matrix (4) is obtained by subjecting the matrix (5) to the similarity transformation[5]

$$\begin{pmatrix} \cos \frac{1}{2}\theta_r & -\sin \frac{1}{2}\theta_r e^{i\phi_r} \\ \sin \frac{1}{2}\theta_r e^{-i\phi_r} & \cos \frac{1}{2}\theta_r \end{pmatrix} \begin{pmatrix} e^{\mu/2} & 0 \\ 0 & e^{-\mu/2} \end{pmatrix} \begin{pmatrix} \cos \frac{1}{2}\theta_r & \sin \frac{1}{2}\theta_r e^{i\phi_r} \\ -\sin \frac{1}{2}\theta_r e^{-i\phi_r} & \cos \frac{1}{2}\theta_r \end{pmatrix}, \quad (8)$$

where θ_r and ϕ_r denote the polar and azimuthal angles of the axis of (4), so that $\sin \theta_r \cos \phi_r = \cos \alpha$, $\sin \theta_r \sin \phi_r = \cos \beta$, and $\theta_r = \gamma$. The matrix on the right

in (8) is the matrix of a rotation that turns the axis of (4) into coincidence with the P_3-axis, and the matrix on the left is the inverse of this rotation. A negative-definite Hermitian transformation can be handled by factoring it into a positive-definite Hermitian transformation and a 360° rotation.[7]

The third type of homogeneous linear transformation that we have to study is the numerical multiple of the identity transformation, whose matrix is the product of the unit matrix by a (complex) number. The determinant of such a transformation is not, in general, unity. The geometrical effect of the transformation is, according to (1), simply to multiply the magnitude of every spinor by a real number and, at the same time, to rotate each spinor about its own axis through some angle, the real number and the angle being the same for all spinors. The direction of a spinor is evidently not changed by this transformation.

As mentioned at the beginning of this section, every non-singular homogeneous linear transformation of the components of a spinor can be uniquely factored into a rotation, a positive-definite unimodular Hermitian transformation, and a multiple of the identity transformation. To prove this, note that any non-singular matrix can be uniquely factored into a unitary matrix and a positive-definite Hermitian matrix,[8] whose determinants will be, respectively a phase factor[9] and a real positive number.[9] Then if we factor out a multiple of the unit matrix whose determinant is equal to that of the original matrix, there will remain a unimodular unitary matrix and a positive-definite unimodular Hermitian matrix.

It follows from this, in particular, that every transformation of the group C_2 can be built up out of one rotation and one unimodular Hermitian transformation. However, there is another way, to be discussed in Sections 7 and 8, of building up a transformation of C_2 out of rotations and unimodular Hermitian transformations, which is better suited to four-terminal network theory.

4. Transformation of the associated vector and of the direction ratio. When the components of a spinor are subjected to a transformation, the components and magnitude of the associated vector must undergo a transformation related in some way to the spinor transformation; and likewise the ratio σ must undergo a related transformation. The nature of the relationships among these transformations will now be studied.

Let the transformation of the components of the spinor be

$$\begin{aligned} S_1' &= a_{11}S_1 + a_{12}S_2 \\ S_2' &= a_{21}S_1 + a_{22}S_2. \end{aligned} \tag{9}$$

Then the transformation equations for the components and magnitude of the associated vector may be written as follows:

$$\begin{pmatrix} P' + P_3' & P_1' + iP_2' \\ P_1' - iP_2' & P' - P_3' \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} P + P_3 & P_1 + iP_2 \\ P_1 - iP_2 & P - P_3 \end{pmatrix} \begin{pmatrix} a_{11}^* & a_{21}^* \\ a_{12}^* & a_{22}^* \end{pmatrix}. \tag{10}$$

[7] See Reference 3, Section 5.

[8] C. C. MacDuffee, *The Theory of Matrices*, Chelsea, New York, 1946, p. 77.

[9] Reference 8, pp. 26, 28.

In this equation the components and magnitude of the vector **P** have been gotten together into a second-rank spinor,[10] and (10) is just the transformation formula for a second-rank spinor. The last matrix on the right in (10) is the Hermitian conjugate of the matrix of the a's. Eq. (10) may be verified most easily by noting that the second-rank spinor is just the direct product[10] of the spinor **S** by itself; that is, the matrix elements of the middle matrix on the right side of (10) are equal, respectively, to $S_1S_1^*$, $S_1S_2^*$, $S_2S_1^*$, and $S_2S_2^*$.

When the spinor components are subjected to the transformation (9), the ratio σ will evidently undergo the linear fractional transformation

$$\sigma' = (a_{11}\sigma + a_{12})/(a_{21}\sigma + a_{22}). \tag{11}$$

It should be noted that the correspondence between the homogeneous linear transformations of a spinor and the linear fractional transformations of σ is not one-to-one, since the coefficients of (11) can all be multiplied by an arbitrary constant without changing the transformation, whereas the multiplying of the coefficients of (9) by an arbitrary factor will yield a different spinor transformation. Thus the group of linear fractional transformations of a complex variable is isomorphic, in one-to-one correspondence, with the group C_2, which is only a subgroup of the entire group of homogeneous linear transformations of two

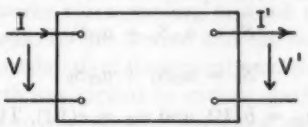


FIG. 4. Four-terminal network

complex variables. This is the reason why the spinor method is applicable to a greater variety of four-terminal networks than the method based on the linear fractional transformations of one complex variable.

5. Application of spinor theory to linear four-terminal networks. Let us consider a linear four-terminal network in the steady state, with input voltage and current V and I and output voltage and current V' and I' (see Figure 4, in which the arrows indicate the positive senses). All four quantities are assumed to be complex and to contain a time-dependent factor $e^{2\pi i\nu t}$; thus

$$V = V_0 e^{i(2\pi\nu t+\delta)} \text{ and } I = I_0 e^{i(2\pi\nu t+\epsilon)}, \tag{12}$$

and similarly for V' and I'. Here δ and ϵ are the initial phases of V and I. Eqs. (12) can be expressed more conveniently in terms of the sum (ψ) of the phases of V and I and the difference (ϕ) of these phases,

$$\psi = 4\pi\nu t + \delta + \epsilon \text{ and } \phi = \delta - \epsilon. \tag{13}$$

Thus,

$$V = V_0 e^{i(\psi+\phi)/2} \text{ and } I = I_0 e^{i(\psi-\phi)/2}. \tag{14}$$

[10] Reference 3, Section 8.

Note the similarity of form between these functions and those in Eqs. (1). Because of this similarity, we may tentatively think of V and I as being the components of a spinor, the "input spinor"; then evidently the azimuthal angle ϕ of this spinor is equal to the angle of lag. The angle ψ, which by (13) is a linear function of the time, specifies the angular position of the spinor about its axis (Figure 1); and therefore the spinor must rotate about its axis with frequency 2ν, that is, in step with the power fluctuations.

Now there is one difficulty in regarding V and I as the components of a spinor: the components of a spinor ought to be dimensionally alike. This difficulty can be surmounted by the use of a dimensional factor, which we shall take to be 1 ohm (1Ω). Then the best way to define the components of the input spinor appears to be as follows:

$$S_1 = V/(1\Omega)^{\frac{1}{2}}, \qquad S_2 = I(1\Omega)^{\frac{1}{2}}. \tag{15}$$

The components S_1' and S_2' of the output spinor are defined similarly.

The network equations, expressed in terms of the voltages and currents, will have the form[11]

$$V' = aV + bI, \qquad I' = cV + dI. \tag{16}$$

In terms of the S's, the network equations will be

$$\begin{aligned} S_1' &= a_{11}S_1 + a_{12}S_2 \\ S_2' &= a_{12}S_1 + a_{22}S_2 , \end{aligned} \tag{17}$$

where $a_{11} = a$, $a_{22} = d$, $a_{12} = b/1\Omega$, and $a_{21} = c(1\Omega)$. The a's are all dimensionless, while the S's all have the dimension (power)$^{\frac{1}{2}}$. The determinant of (17), $a_{11}a_{22} - a_{12}a_{21}$, is equal to the determinant of (16), $ad - bc$. This determinant is not necessarily equal to 1, since the circuit may contain amplifiers.

If voltage and current are essentially the components of a spinor, then the components of the associated vector ought to have some meaning. These components can be obtained, by using Eqs. (2) and (15), in terms of voltage and current; the results at the input side are

$$\begin{aligned} P_1 &= \mid VI \mid \cos\phi = 2 \times \text{average power used} \\ P_2 &= \mid VI \mid \sin\phi = 2 \times \text{reactive power} \\ P_3 &= \tfrac{1}{2}[(VV^*/1\Omega) - (II^*)1\Omega] \\ P &= \tfrac{1}{2}[(VV^*/1\Omega) + (II^*)1\Omega]. \end{aligned} \tag{18}$$

The component P_3 may be called the power difference, for it is equal to the difference between the power that the voltage V would supply to a resistance of 1 ohm and the power that the current I would dissipate in a resistance of 1 ohm. The value of P_3 is a measure of whether, and to what extent, the power is being transmitted at high voltage and low current or the opposite.

[11] See any of the standard books on communications circuits, or see Brillouin, *l.c.* Chapter 9.

It appears at first sight that P_1 must always be positive. However, P_1 may be considered to have a negative value if the side of the four-terminal network that is nearer to the generator is called the output, rather than the input, side. In this case, the average "output" power P_1' from a passive network will be more negative than the average "input" power P_1, and V and I will be more than 90° out of phase, according to the conventions of sense shown in Figure 4.

Because of Eqs. (15), the ratio σ, already discussed in Section 2, is evidently equal to the input impedance of the network divided by 1 ohm. The values of σ at the different points on the surface of the unit sphere (Figure 2) can thus be thought of as impedance values, the points on the equator being those whose impedance has an absolute value of 1 ohm (since at the equator $\cot \frac{1}{2}\theta = 1$). This plotting of impedance values on the unit sphere can also be accomplished by first plotting impedance values according to the usual rules in the P_1P_2-plane and then transferring them by stereographic projection to the surface of the sphere, as illustrated by Figure 2.

The output power components can be expressed in terms of the input power components by the transformation (10); and the output impedance is related to the input impedance by a linear fractional transformation that is obtained by multiplying (11) by 1 ohm.

6. Four-terminal networks corresponding to some rotations and unimodular Hermitian transformations. In this section and the next two sections, we shall undertake to show how the details of the general geometric scheme work out when the rules already set forth are applied to various special cases. First let us find networks corresponding to the simplest rotations and unimodular Hermitian transformations.

The unimodular Hermitian transformations having for their axes the P_1, P_2, and P_3 axes are respectively, by (4):

$$\begin{pmatrix} \cosh \frac{1}{2}\mu & \sinh \frac{1}{2}\mu \\ \sinh \frac{1}{2}\mu & \cosh \frac{1}{2}\mu \end{pmatrix}, \quad \begin{pmatrix} \cosh \frac{1}{2}\mu & i \sinh \frac{1}{2}\mu \\ -i \sinh \frac{1}{2}\mu & \cosh \frac{1}{2}\mu \end{pmatrix}, \quad \begin{pmatrix} e^{\mu/2} & 0 \\ 0 & e^{-\mu/2} \end{pmatrix}. \tag{19}$$

The third of these is the easiest to deal with. It is the matrix of an ideal transformer, as is seen by comparing it with Eqs. (16) and (17). The voltage ratio is $e^{\mu/2}$, and the voltage is stepped up or down according as μ is positive or negative. From the geometrical picture, Figure 3a, it is evident that P_1, P_2, ϕ, and ψ are invariant, and therefore V' is in phase with V and I' with I.

The first and second of the matrices (19) have their diagonal elements equal, and the corresponding networks must therefore be symmetric networks.[11] For the first matrix the corresponding network can be any symmetric, purely resistive network (e.g. a T-network) whose circuit elements have been so chosen that $(b/c)^{\frac{1}{2}}$ shall be equal to 1 ohm, for then a_{12} will be equal to a_{21} and therefore the matrix will have the form given in (19). If the network resistances are all positive, then μ will be negative, and P_1' will be less than P_1 for every spinor. If the resistances are all negative, μ will be positive and P_1' will be greater than P_1. The condition that $(b/c)^{\frac{1}{2}}$ shall be 1 ohm is equivalent to saying that the itera-

tive (or image) impedance[11] of the network shall be 1 ohm. The propagation (or transfer) constant ($\frac{1}{2}\mu$) is real.

For the network corresponding to the second matrix of (19), we can take any symmetric, purely reactive network such that b and c are imaginary and of opposite sign and $| b/c |^{\frac{1}{2}}$ is equal to 1 ohm (e.g. a symmetric network containing only inductive reactance or else only capacitive reactance). These conditions are equivalent to saying that the iterative impedances shall be $\pm i$ ohms and the propagation constant real. If all the reactances in the network are inductive, μ will be negative; if capacitive, μ will be positive. It is evident from the picture of a Hermitian transformation (Figure 3a, but with the spheroid's major axis along the P_2-axis) that, for every spinor, P_1 and P_3 are invariant to the transformation. However, P_2 is changed, and so is ϕ unless ϕ is $\pm 90°$.

A unimodular Hermitian transformation whose axis lies in the P_1P_2-plane and makes an arbitrary angle α with the P_1-axis has its diagonal coefficients both equal to $\cosh \frac{1}{2}\mu$ and its off-diagonal coefficients equal to $\sinh \frac{1}{2}\mu e^{\pm i\alpha}$, as follows from (4). For the corresponding network, one can take a symmetric network whose elements are impedances with phase angle α and absolute values so chosen that $| b/c |^{\frac{1}{2}} = 1$ ohm. The iterative impedance is $e^{i\alpha}$ ohms and the propagation constant is again real and equal to $\frac{1}{2}\mu$.

The matrices of rotations about the P_1, P_2, and P_3 axes are, by (3):

$$\begin{pmatrix} \cos \frac{1}{2}\kappa & i \sin \frac{1}{2}\kappa \\ i \sin \frac{1}{2}\kappa & \cos \frac{1}{2}\kappa \end{pmatrix}, \quad \begin{pmatrix} \cos \frac{1}{2}\kappa & -\sin \frac{1}{2}\kappa \\ \sin \frac{1}{2}\kappa & \cos \frac{1}{2}\kappa \end{pmatrix}, \quad \begin{pmatrix} e^{i\kappa/2} & 0 \\ 0 & e^{-i\kappa/2} \end{pmatrix}. \tag{20}$$

A rotation about the P_1-axis can be realized by a symmetric lossless filter section, operating in its passband and having its circuit elements so chosen that $| b/c |^{\frac{1}{2}} = 1$ ohm; such a network will have its iterative impedance equal to 1 ohm and its propagation constant imaginary ($\frac{1}{2}i\kappa$). The power component P_1 is invariant, as is the power magnitude P; while P_2 and P_3 undergo an orthogonal transformation. If the filter section is operating outside its passband and has $| b/c |^{\frac{1}{2}}$ equal to 1 ohm, the corresponding transformation will be a Hermitian transformation whose axis is the P_2-axis.

A rotation about the P_2-axis can be realized by a symmetric combination of positive and negative resistances, arranged in the circuit in the same way in which the positive and negative reactances are arranged in a symmetric filter section, with the values so chosen that the iterative impedances will be $\pm i$. Similarly, a rotation about any arbitrary axis in the P_1P_2-plane can be realized by a symmetric combination of positive and negative impedances of suitably chosen phase angle, just as was done in the case of Hermitian transformations.

The rotation about the P_3-axis can be realized by a combination of networks of the types just discussed, since any rotation can be built up out of rotations about two arbitrarily chosen perpendicular axes.[12]

7. Geometry of transformations corresponding to arbitrary symmetric four-terminal networks with determinant 1. A symmetric network has a equal to

[12] Reference 3, Section 4.

d; and let us assume for the present that neither b nor c vanishes. Then the problem can be referred to the cases studied in the preceding section, in the following way. The matrix of the a's, in Eq. (17), can be derived from another matrix that has its off-diagonal elements, b_{12} and b_{21}, equal in absolute value, by the similarity transformation,

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = \begin{pmatrix} |\, a_{12}/a_{21}\,|^{\frac{1}{2}} & 0 \\ 0 & |\, a_{21}/a_{12}\,|^{\frac{1}{2}} \end{pmatrix} \begin{pmatrix} a_{11} & b_{12} \\ b_{21} & a_{11} \end{pmatrix} \begin{pmatrix} |\, a_{21}/a_{12}\,|^{\frac{1}{2}} & 0 \\ 0 & |\, a_{12}/a_{21}\,|^{\frac{1}{2}} \end{pmatrix}, \quad (21)$$

where

$$\begin{aligned} b_{12} &= |\, a_{12}a_{21}\,|^{\frac{1}{2}} \times \text{phase factor of } a_{12} \\ b_{21} &= |\, a_{12}a_{21}\,|^{\frac{1}{2}} \times \text{phase factor of } a_{21}. \end{aligned} \quad (22)$$

Since $|\, b_{12}\,| = |\, b_{21}\,|$, the matrix containing the b's represents a rotation or a Hermitian transformation (or a coaxial combination of these) whose axis is in the P_1P_2-plane.

The geometrical interpretation of (21) is as follows. The first matrix on the left of the righthand side of (21) represents a Hermitian transformation with axis along the P_3-axis; therefore it transforms any sphere into a prolate spheroid, as depicted in Figure 3a. Let us start with such a spheroid and apply the entire right side of (21) to spinors terminating on it; this will first transform the spheroid into a sphere, then subject all the spinors to a rotation or Hermitian transformation (or coaxial combination) about an axis in the P_1P_2-plane, and finally transform the sphere back into a spheroid. The last one of these three transformations converts the axis of the rotation or Hermitian transformation (or combination) into a bent line (a trace of the cone shown in Figure 3b); the two parts of this line are the principal axes of the entire transformation (21), because a spinor lying along one of them does not have its direction changed but only its magnitude and/or its angular position about its axis. The P_3-axis bisects the angle between these principal axes; this is evidently the geometric expression of the fact that the network is symmetric. It follows from Section 3 and Eq. (21) that the cotangent of half the angle between either principal axis and the P_3-axis is equal to $|\, a_{12}/a_{21}\,|^{\frac{1}{2}}$ or, in terms of the original network coefficients, to $|\, b/c\,|^{\frac{1}{2}}(1\Omega)^{-1}$.

Since a spinor lying along either principal axis of the transformation (21) has only its magnitude and/or its angle ψ changed, the effect of (21) on such a spinor will be merely to multiply its components by a (complex) numerical factor. The two factors associated with the two principal axes are the principal values of the transformation (21). They are reciprocals of each other, since the determinant of (21) is 1. If the middle transformation on the right side of (21) is a rotation, the principal values will be phase factors whose phases are plus or minus half the rotation angle; if it is a Hermitian transformation, they will be real numbers whose logarithms are plus or minus half the parameter of the Hermitian transformation. The propagation constant of the network, which is by definition the logarithm of one of the principal values, will be equal to $\frac{1}{2}i$ times

the rotation angle, or half the parameter of the Hermitian transformation, as the case may be. In the case of a combination of a rotation and a Hermitian transformation the propagation constant will, of course, be complex.

It is not difficult to visualize the effect of the transformation (21) as a whole on spinors that do not lie along a principal axis. If the middle transformation on the right side of (21) is a rotation, then the blade of each spinor will move along an ellipse which is the intersection of the prolate spheroid of Figure 3a with a vertical plane. If the middle transformation is a Hermitian transformation, each spinor blade will move along a hyperbola that is the intersection of one of the hyperboloids of Figure 3b with a vertical plane.

Let us next consider the case where the network coefficient c vanishes (e.g. a network containing series impedance but no shunt elements). This is the limiting case obtained from the type of networks just studied as the angle between either principal axis and the P_3-axis is decreased to zero, because, as already shown, the cotangent of half this angle is equal to $|\, b/c \,|^{\frac{1}{2}}(1\Omega)^{-1}$. If c vanishes, therefore, the principal axes both coincide with the positive P_3-axis. Similarly, if $c \neq 0$ but $b = 0$, the principal axes coincide with the negative P_3-axis; this case is realized by a network having shunt impedance only.

8. Geometry of transformations corresponding to arbitrary four-terminal networks of determinant 1. The principal values λ (or characteristic roots, as they are sometimes called) of the general transformation (17) and the σ-values of its principal axes can be found by the well-known method of solving the secular (or characteristic) equation.[5] The results are:

$$\begin{aligned} \lambda &= \tfrac{1}{2}\{a_{11} + a_{22} \pm [(a_{11} + a_{22})^2 - 4]^{\frac{1}{2}}\} \\ \sigma &= (1/2a_{21})\{a_{11} - a_{22} \pm [(a_{11} + a_{22})^2 - 4]^{\frac{1}{2}}\}. \end{aligned} \tag{23}$$

Since the determinant of the transformation is assumed to be 1, the two values of λ are reciprocal. From the second of Eqs. (23), it is seen that the two σ values will have their signs changed if a_{11} and a_{22} are interchanged; or, in other words, if the input connections of the network are interchanged with the output connections, the new principal axes will be the images, in the P_3-axis, of the original principal axes. If the four values of σ thus obtained for both sets of axes are multiplied by 1 ohm, the resulting four quantities will be the iterative impedances of the network[12]; this follows at once from a comparison of the definitions of principal axis and iterative impedance. Most authors discard the two iterative impedances that have a negative real part. In the case of a symmetric network, two of the iterative impedances coincide with the other two, since the two σ-values in (23) are in that case the negatives of each other.

Now the two principal axes of the general unimodular homogeneous linear transformation are geometrically related to the bisector of their angle in the same way in which, in the preceding section, the principal axes of a transformation having $a_{11} = a_{22}$ were related to the P_3-axis. Therefore it must be possible to use the same method as in the preceding section; that is, it must be possible

[12] The relation between iterative impedances and principal axes has been noted by Brillouin, *l.c.* Section 54.

to derive the matrix of the general transformation, by means of a similarity transformation, from another matrix that will be either unitary or Hermitian or a coaxial combination of the two. The transforming matrix in the similarity transformation (i.e. the matrix corresponding to the righthand matrix in (21)) must now be a unimodular Hermitian matrix whose axis is the bisector of the angle between the principal axes of the given transformation, and the parameter of this matrix must be such that the first step of the similarity transformation will cause the principal axes of the given transformation to become antiparallel to each other. This is exactly what was done in the preceding section. If the second matrix of the similarity transformation is unitary, then the principal values of the original transformation are reciprocal phase factors whose phases are plus or minus half the rotation angle κ; the propagation constant is then $\frac{1}{2}i\kappa$. If the second matrix is Hermitian, the principal values are reciprocal real numbers whose logarithms are plus or minus half the parameter of the Hermitian transformation (i.e. $\pm \frac{1}{2}\mu$); the propagation constant is $\frac{1}{2}\mu$. All these conclusions follow by analogy with the preceding section. Argument by analogy is justified here because the spinor properties concerned depend only on the geometrical

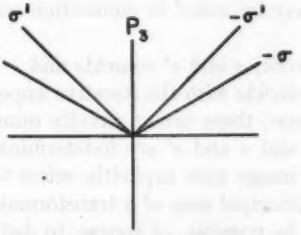


FIG. 5. Image axes

relationships among the various axes involved and not on their relation to the P_3-axis.

Thus, in the spinor scheme, the iterative propagation constant of a network appears as a rotation angle (times $\frac{1}{2}i$), or as a real exponent, or as a complex linear combination of the two. To these three cases there correspond, in the scheme based on the linear fractional transformations of the complex variable σ (Section 4), the so-called elliptic, hyperbolic, and loxodromic transformations respectively.[14] If the two principal axes of the network transformation coincide, the corresponding σ transformation is parabolic.[14]

9. Image axes of a transformation and image impedances. It has already been noted (Section 2) that any two directions whose σ-values are the negatives of each other are the images of each other in the P_3-axis, and also that in the case of a symmetric network the principal axes satisfy this relation. In the case of a general network, it is possible to find two directions that are the images of each other in the P_3-axis and are transformed by the network equations into two directions that are likewise the images of each other in the P_3-axis. (See Figure 5, but

[14] See van Slooten, *l.c.*, or any standard book on functions of a complex variable.

note that σ' and $-\sigma'$ are not necessarily coplanar with σ and $-\sigma$.) The four directions may be called image axes, and the four σ values times 1 ohm are the image impedances of the network, the two with negative real parts being discarded as meaningless by most authors.

It is not difficult to find the σ values for the image axes in terms of the coefficients of the network equations. The transformation equations for σ and $-\sigma$ are, by (11):

$$\sigma' = (a_{11}\sigma + a_{12})/(a_{21}\sigma + a_{22})$$
$$-\sigma' = (-a_{11}\sigma + a_{12})/(-a_{21}\sigma + a_{22}). \tag{24}$$

Eliminate σ', and solve for σ and then for σ'. The results are:

$$\sigma = (a_{12}a_{22}/a_{11}a_{21})^{\frac{1}{2}} \qquad \sigma' = (a_{11}a_{12}/a_{21}a_{22})^{\frac{1}{2}}. \tag{25}$$

On multiplying these values by ± 1 ohm, one gets the standard formulas for the image impedances,

$$\pm(bd/ac)^{\frac{1}{2}} \text{ and } \pm(ab/cd)^{\frac{1}{2}}. \tag{26}$$

The following facts may be noted in connection with image axes and image impedances:

For a symmetric network, σ and σ' coincide and $-\sigma$ and $-\sigma'$ coincide; also, the image impedances coincide with the iterative impedances.

For an ideal transformer, there are an infinite number of sets of image axes and image impedances, and σ and σ' are indeterminate.

The definition of the image axes explicitly refers to the P_3-axis, in contrast to the definition of the principal axes of a transformation, which is independent of the P-axes. It would be possible, of course, to define image axes referred to any arbitrarily chosen direction of the $P_1P_2P_3$-space.

If the network transformation is a pure rotation, the image axes are in the P_1P_2-plane and the values of σ and σ' are $ie^{i\phi_r}$ and $ie^{-i\psi_r}$, where ϕ_r and ψ_r are the second and third Eulerian angles of the rotation.[12] (The image transfer constant in this case is $\frac{1}{2}i$ times the first Eulerian angle of the rotation.)

10. Negative resistances and amplifier circuits. It has been assumed up to this point that the four-terminal network could contain negative resistances. If these are ruled out, the question arises which of the various spinor transformations will be discarded. Since the network is now purely passive, all transformations will be disallowed that cause P_1' to exceed P_1 for one or more of all spinors. For example, all rotations will be ruled out except rotations about the P_1-axis; and every Hermitian transformation will be ruled out whose spheroid has its center in front of the P_2P_3-plane. The remaining transformations do not, of course, form a group, because most of them no longer have inverses.

A linear amplifier circuit obeys Eqs. (16) or (17), but with the determinant not equal to 1. As noted at the end of Section 3, the matrix of the equations can always be factored into a matrix of determinant 1 and a numerical multiple of

the unit matrix. The first of these factors corresponds to a network that satisfies the reciprocity theorem, and can be treated according to the rules already developed in Sections 6, 7, and 8. The numerical multiple of the unit matrix has a simple geometrical meaning already noted in Section 3: it multiplies the amplitudes of voltage and current by a real number and/or shifts the phases of voltage and current by some angle, the real number and the phase-shift angle being the same for all spinors. Since this type of transformation is what distinguishes a linear amplifier circuit from a circuit satisfying the reciprocity theorem, a hypothetical circuit corresponding to this type of transformation might be called an "ideal amplifier." The ideal amplifier differs from the ideal transformer in that it multiplies both current and voltage by the same factor, whereas an ideal transformer multiplies them by reciprocal factors.

MICHIGAN STATE COLLEGE

(Received June 30, 1952)

SECONDARY FLOW WITHIN A BEND*

By Hans P. Eichenberger

1. Introduction. In analysing flow in fluid machines, one usually assumes the existence of potential flow. This assumption implies that the stagnation pressure is the same on different streamlines, and simplifies the calculation of the flow phenomena considerably. In the case of incompressible flow, the problem reduces to the solution of the Laplace equation. But in many machines the stagnation pressure is not constant throughout. For instance, in an axial flow compressor which is fed from the free atmosphere the flow at the inlet is, except in a very thin boundary layer, potential flow. However, after a few stages of a multistage compressor the stagnation pressure is only nearly invariant going from streamline to streamline (1). It varies, however, in direction of the span of the blades only (from hub to casing of the compressor) and remains essentially constant in direction perpendicular to the span. The flow which occurs under these circumstances deviates from potential flow. The difference between the real flow and the potential flow has been called secondary flow.

Until recently the effects of this secondary flow were neglected in the design of axial flow compressors. It is clear that as a result the blades may operate far from the expected design conditions and separation at the root and tip of a blade may occur due to this incorrect design. This phenomenon contributes to the secondary losses which account for approximately half of the losses occurring in a compressor (7). Studies conducted at the Gas Turbine Laboratory at M.I.T. (8) seem to indicate that at least a part of these losses could be avoided if account were taken of the secondary flow in the design of the cascade instead of assuming the existence of potential flow as was done in the past.

The proposed method is to add a correction to the blades designed on the basis of potential flow. In order to calculate the correction, knowledge of the secondary flow is necessary. In this paper it is shown how secondary flow can be calculated.

Although the fact that stagnation pressure varies within the stream shows that dissipation occurred, we will assume in this paper an inviscid fluid. This assumption is justified by experiments (8) which showed that the effect of viscosity within one passage is negligible although the cumulative effect over many stages of a compressor leads to a flow with vorticity.

2. The Problem. The rather complicated geometry of a blade passage in a compressor was replaced by a simple bend such that the experimental result could easily be interpreted in the light of the theory. The surface of the blades was replaced by two coaxial cylindrical surfaces bounded by two horizontal planes which are a distance $2h$ apart as shown on Figure 1.

* This work has been sponsored by the ONR under contract N5ori 07848 and was performed in the Gas Turbine Lab. at M.I.T. under the direction of E. S. Taylor. The author is also indebted to Prof. C. C. Lin for his suggestions and interest in this paper.

The problem can be formulated as follows: Given at a certain section, say $\varphi = 0$, of this bend $W = 0$, $U = 0$ and $V(r, z)$ in such a fashion that the stagnation pressure varies in z direction but not in r direction. U, V, W are the velocities in direction of r, φ, and z, the three orthogonal coordinates as indicated on Fig. 1.

We wish to find the velocities downstream at a bend section which corresponds to a turning angle met with in compressors. The calculation of these velocities form the main part of this paper.

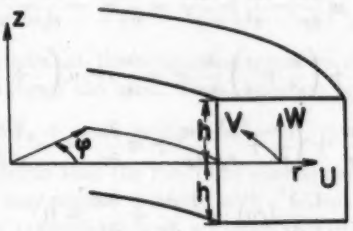


FIG. 1

The flow is governed by the three momentum equations and the continuity equation which are, as written in cylindrical coordinates:

$$U\frac{\partial U}{\partial r} + V\frac{1}{r}\frac{\partial U}{\partial \varphi} + W\frac{\partial U}{\partial z} - \frac{V^2}{r} = -\frac{1}{\rho}\frac{\partial P}{\partial r} \tag{A}$$

$$U\frac{\partial V}{\partial r} + V\frac{1}{r}\frac{\partial V}{\partial \varphi} + W\frac{\partial V}{\partial z} + \frac{UV}{r} = -\frac{1}{\rho r}\frac{\partial P}{\partial \varphi} \tag{B}$$

$$U\frac{\partial W}{\partial r} + V\frac{1}{r}\frac{\partial W}{\partial \varphi} + W\frac{\partial W}{\partial z} = -\frac{1}{\rho}\frac{\partial P}{\partial z} \tag{C}$$

$$\frac{1}{r}\frac{\partial}{\partial r}(rU) + \frac{1}{r}\frac{\partial V}{\partial \varphi} + \frac{\partial W}{\partial z} = 0 \tag{D}$$

where P is the pressure and ρ is the density.

3. The Linearization. A first approximation to the above problem is the well-known potential flow around the bend of Figure 1. After a very short bend angle this potential flow is very nearly equal to the free vortex flow. This flow is defined by $U = 0$, $V = c/r$, $W = 0$, where c is a constant depending on the mass rate of flow. If we put these velocities into the equations (A) to (C), we obtain for the corresponding pressure, which we call the pressure of the potential flow P_p:

$$\frac{\partial P_p}{\partial \varphi} = 0 \qquad \frac{\partial P_p}{\partial z} = 0 \qquad \frac{\partial P_p}{\partial r} = \rho\frac{V^2}{r} = \rho\frac{c^2}{r^3} \tag{E}$$

The index p denotes the value of P in the case of potential flow.

We assume now that the real flow deviates only little from the above defined free vortex flow such that we can write:

$$U = 0 + u(r, \varphi, z), \qquad V = (c/r) + v(r, \varphi, z),$$
$$W = 0 + w(r, \varphi, z), \qquad P = P_p + p(r, \varphi, z)$$

where u, v, w are small as compared to c/r and p small as compared to P_p.

If we put these expressions into equations A to D we obtain:

$$u\frac{\partial u}{\partial r} + \left(\frac{c}{r^2} + \frac{v}{r}\right)\frac{\partial u}{\partial \varphi} + w\frac{\partial u}{\partial z} - \frac{c^2}{r^3} - \frac{2cv}{r^2} - \frac{v^2}{r} = -\frac{c^2}{r^3} - \frac{1}{\rho}\frac{\partial P}{\partial r}$$

$$-u\frac{c}{r^2} + u\frac{\partial v}{\partial r} + \left(\frac{c}{r^2} + \frac{v}{r}\right)\frac{\partial v}{\partial \varphi} + w\frac{\partial v}{\partial z} + u\left(\frac{v}{r} + \frac{c}{r^2}\right) = \frac{-1}{\rho r}\frac{\partial P}{\partial \varphi}$$

$$u\frac{\partial w}{\partial r} + \left(\frac{c}{r^2} + \frac{v}{r}\right)\frac{\partial w}{\partial \varphi} + w\frac{\partial w}{\partial z} = -\frac{1}{\rho}\frac{\partial P}{\partial z}$$

$$\frac{1}{r}\frac{\partial}{\partial r}(ru) + \frac{1}{r}\frac{\partial v}{\partial \varphi} + \frac{\partial w}{\partial z} = 0 \tag{4}$$

If we assume that $\partial u/\partial r$, $r^{-1}\partial u/\partial \varphi$, $\partial u/\partial z$ are of the same order and $u, v, w, \ll c/r$, considering relation (E), the above equations reduce to:

$$\frac{c}{r^2}\frac{\partial u}{\partial \varphi} - \frac{2cv}{r^2} = -\frac{1}{\rho}\frac{\partial p}{\partial r}, \tag{1}$$

$$\frac{c}{r^2}\frac{\partial v}{\partial \varphi} = -\frac{1}{\rho r}\frac{\partial p}{\partial \varphi}, \tag{2}$$

$$\frac{c}{r^2}\frac{\partial w}{\partial \varphi} = -\frac{1}{c\rho}\frac{\partial p}{\partial z} \tag{3}$$

We now have the task of finding a solution to the four linear partial differential equations (1) to (4) for the four unknowns u, v, w, and p, all functions of the three independent variables r, φ and z. As boundary conditions we have, of course, the solid walls of the bend and we wish to impose $u = w = 0$ at the bend section $\varphi = 0$. In addition, we specify $v(r, z)$ at this section.

4. Derivation of the Partial Differential Equation. Integration of (2) with respect to φ yields immediately:

$$(c/r^2)v + (1/\rho r)p = f(r, z) \tag{2'}$$

$f(r, z)$ standing for an arbitrary function $(c/r^2)v_0 + (1/\rho r)p_0 = f(r, z)$ where the index $_0$ denotes the values of the functions at $\varphi = 0$. Taking the derivative of (2') with respect to z and r gives, respectively:

$$\frac{c}{r^2}\frac{\partial v}{\partial z} + \frac{1}{\rho r}\frac{\partial p}{\partial z} = \frac{\partial}{\partial z}\left(\frac{c}{r^2}v_0 + \frac{1}{\rho r}p_0\right) \tag{5}$$

and

$$\frac{1}{\rho}\frac{\partial p}{\partial r} = -\frac{\partial}{\partial r}\left(\frac{cv}{r}\right) + \frac{\partial}{\partial r}\left(\frac{cv_0}{r} + \frac{1}{\rho}p_0\right) \tag{6}$$

From (6) and (1) we can eliminate $\partial p/\partial r$:

$$\frac{c}{r^2}\frac{\partial u}{\partial \varphi} - \frac{2cv}{r^2} + \frac{c}{r^2}v - \frac{1}{r}\frac{\partial v}{\partial r} = -\frac{\partial}{\partial r}\left(\frac{cv_0}{r} + \frac{1}{\rho}p_0\right)$$

$$\frac{c}{r^2}\frac{\partial u}{\partial \varphi} - \frac{c}{r^2}\frac{\partial}{\partial r}(rv) = -\frac{\partial}{\partial r}\left(\frac{cv_0}{r} + \frac{1}{\rho}p_0\right) \tag{7}$$

and with (3) eliminating $\partial p/\partial z$ from (5)

$$\frac{c}{r^2}\frac{\partial w}{\partial \varphi} - \frac{c}{r}\frac{\partial v}{\partial z} = -\frac{\partial}{\partial z}\left(\frac{cv_0}{r} + \frac{1}{\rho}p_0\right) \tag{8}$$

In section 2 we stated that the stagnation pressure does not vary in radial direction at the beginning of the bend. The stagnation pressure is by definition:

$$S = P + \tfrac{1}{2}\rho V^2 = P_p + p + \tfrac{1}{2}\rho(V_p + v)^2 = P_p + p + \tfrac{1}{2}\rho[(c/r) + v]^2,$$

where the indices p denote that the functions correspond to the potential flow. In this expression, we may neglect the term with v^2. Also, we know that the stagnation pressure of the potential flow is invariant; so that,

$$(\partial/\partial r)(P_p + \tfrac{1}{2}\rho V_p^2) = 0$$

Thus, at the beginning of the bend, from $\partial S/\partial r = 0$ follows

$$(\partial/\partial r)[\rho v_0(c/r) + p_0] = 0$$

which means that the right hand member of (7) is zero.

In this case (7) is satisfied if we define a function Φ:

$$r^{-1}\partial\Phi/\partial\varphi = v \qquad \partial\Phi/\partial r = u$$

As Φ is a function of the three variables, equation (9) does not define Φ completely. In addition, we may specify its value along a vertical line which will be done later at our convenience.

Writing (4) and (8) in terms of Φ:

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial\Phi}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2\Phi}{\partial\varphi^2} + \frac{\partial w}{\partial z} = 0 \tag{10}$$

$$\frac{\partial w}{\partial\varphi} - \frac{\partial^2\Phi}{\partial\varphi\partial z} + r^2\frac{\partial}{\partial z}\left(\frac{v_0}{r} + \frac{p_0}{\rho c}\right) = 0 \tag{11}$$

Integrating (11) with respect to φ:

$$w - \frac{\partial\Phi}{\partial z} + r^2\frac{\partial}{\partial z}\left(\frac{v_0}{r} + \frac{p_0}{\rho c}\right)\varphi = f(r, z) \tag{11a}$$

In order to determine the arbitrary function $f(r, z)$ we use the boundary condition at $\varphi = 0$ namely $w = 0$. Equation (11a) then becomes:

$$\partial\Phi(r, 0, z)/\partial z = -f(r, z)$$

Putting this value of $f(r, z)$ into (11a) yields after differentiation with respect to z:

$$\frac{\partial w}{\partial z} - \frac{\partial^2 \Phi}{\partial z^2} + r^2 \varphi \frac{\partial^2}{\partial z^2}\left(\frac{v_0}{r} + \frac{p_0}{\rho c}\right) = -\frac{\partial^2 \Phi(r, 0, z)}{\partial z^2} \tag{12}$$

eliminating $\partial w/\partial z$ from (10) and (12) we have

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial \Phi}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2 \Phi}{\partial \varphi^2} + \frac{\partial^2 \Phi}{\partial z^2} - \frac{\partial^2 \Phi(r, 0, z)}{\partial z^2} = \varphi r^2 \frac{\partial^2}{\partial z^2}\left(\frac{v_0}{r} + \frac{p_0}{\rho c}\right)$$

In order to determine Φ completely we assume $\Phi = 0$ on the line $\varphi = 0$, $r = r_0$, (r_0 is the radius r at the outside of the bend) because this simplifies the above equation considerably. As $u = 0$ at $\varphi = 0$: $\partial\Phi/\partial r = 0$ there and $\Phi = 0$ in the whole $\varphi = 0$ plane. Therefore,

$$\partial^2 \Phi(r, 0, z)/\partial z^2 \doteq 0 \tag{13}$$

and the partial differential equation to solve is

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial \Phi}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2 \Phi}{\partial \varphi^2} + \frac{\partial^2 \Phi}{\partial z^2} = \varphi r^2 \frac{\partial^2}{\partial z^2}\left(\frac{v_0}{r} + \frac{p_0}{\rho c}\right) \tag{14}$$

The boundary conditions are the following:

(1) $\Phi(r, 0, z) = 0$ was found above and implies that $u = w = 0$ at the beginning of the bend.

(2) $w(r, \varphi, \pm h) = 0$ yields with (11a)

$$\frac{\partial \Phi(r, \varphi, \pm h)}{\partial z} = \varphi r^2 \frac{\partial}{\partial z}\left(\frac{v_0}{r} + \frac{p_0}{\rho c}\right)$$

(3) Finally the condition of no flow across the cylindrical walls can be expressed as

$$\partial\Phi(r_0, \varphi, z)/\partial r = 0 \qquad \partial\Phi(r_i, \varphi, z)/\partial r = 0$$

where r_i is the radius r at the inside of the bend.

The question which could not be answered at the end of last paragraph as to the necessary and sufficient conditions to impose at the bend section $\varphi = 0$ is now solved: we can arbitrarily prescribe v and p, called p_0 and v_0, there, but these functions of r and z always appear in the differential equation (14) and in the boundary conditions in the combination $(\partial/\partial z)[(v_0/r) + (p_0/\rho c)]$, which was shown to be the variation of the stagnation pressure. Thus not the distribution of v or p upstream of the bend, but the variation of the stagnation pressure is responsible for the secondary flow. We should give Φ or its normal derivative on a closed boundary because of the elliptic form of the partial differential equation. But we do not know what conditions we should prescribe downstream; the flow there is precisely what we intend to calculate. We choose the following artifice. We prolong the bend up to 360°. Then we hope that the conditions imposed there do not appreciably affect the flow in the first quarter of the bend.

This assumption is similar to the principle of Saint Venant in elasticity where the equations are of the elliptic type too. Thus we assume arbitrarily

$$\Phi(r, 2\pi, z) = 0$$

which determines the problem completely.

5. Discussion. The components of the vorticity

$$\xi = \frac{1}{r}\frac{\partial w}{\partial \varphi} - \frac{\partial v}{\partial z}, \qquad \eta = \frac{\partial u}{\partial z} - \frac{\partial w}{\partial r}, \qquad \zeta = \frac{1}{r}\frac{\partial}{\partial r}(rv) - \frac{\partial u}{\partial \varphi}$$

in direction of r, φ and z respectively, follow from equations (7), (8) and (11a):

$$\xi = -r\frac{\partial}{\partial z}\left(\frac{v_0}{r} + \frac{p_0}{\rho c}\right), \qquad \eta = 2r\varphi\frac{\partial}{\partial z}\left(\frac{v_0}{r} + \frac{p_0}{\rho c}\right), \qquad \zeta = r\frac{\partial}{\partial r}\left(\frac{v_0}{r} + \frac{p_0}{\rho c}\right) \tag{15}$$

It follows from the linearization that ξ and ζ do not change with bend angle ($\zeta = 0$ in our case) whereas η increases linearly with φ. This has already been found by Squire and Winter and they used this knowledge to calculate the corresponding flow in a straight duct assuming two-dimensional flow, i.e., no change in direction of the main flow. From the linear increase of η, it follows that no periodic phenomenon is possible. If the equations are not linearized the phenomenon of secondary flow appears to be periodic (9). The period, however, is too large to be of any significance for this problem. In other words, the linearization leads to significant results in this case, as will be shown by the experiments.

A comparison of the results of Squire and Winter's method and of the above solution is of interest.

6. The Solution of the Partial Differential Equation. In general it is best to solve (14) by relaxation. If the stagnation pressure varies linearly, however, equation (14) reduces to Laplace's equation and an analytic solution is useful. By separation of variables one obtains the solution

$$\Phi = -\sum_m \sum_k (-1)^k \frac{2s}{mk} \cdot \frac{\sinh mz}{\cosh mh} \cdot \sin k\varphi \cdot c(m, k) \cdot B_k(mr) \tag{15}$$

where

$$s = (\partial/\partial z)[(v_0/r) + (p_0/\rho c)]$$

$B_k(mr)$ is defined for $k = 1, 2, 3 \cdots$ by

$$r^2B''(r) + rB'(r) + [(mr)^2 - k^2]B(r) = 0$$

$$B'(r_i) = 0 \qquad B'(r_0) = 0 \tag{15a}$$

m are the eigenvalues and $c(m, k)$ is determined by the condition:

$$\sum m\, c(m, k) B_k(mr) = r^2 \qquad \text{for all } k. \tag{16}$$

Instead of using the tabulated Bessel function it proved to be faster to approximate each B_k by two trigonometric functions determining the constant by collocation (3).

7. Comparison of Result with Experiment and Squire's Solution. The experiment was performed with a 90° elbow with 8 x 8 inches square cross section and 28 inches mean radius. Upstream of the elbow a stagnation pressure profile as shown on Figure 2.1 was produced. The stagnation pressure was, as nearly as possible, made to increase linearly from the top to the middle and from the bottom to the middle of the bend. Because this was not exactly true, the accuracy of the result is only about 5 per cent.

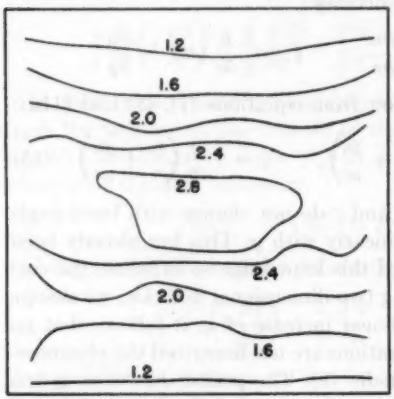


FIG. 2.1

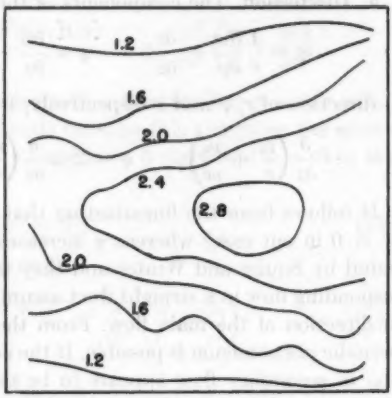


FIG. 2.2

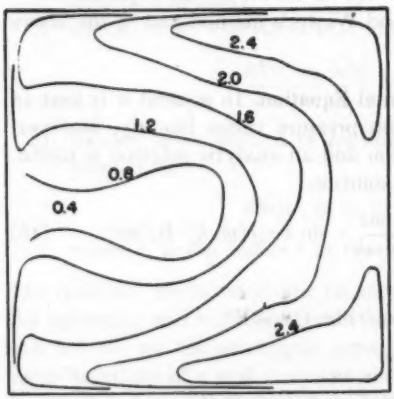


FIG. 2.3

FIG. 2.1–2.3. Map of stagnation pressures in inches water at 0°, 30° and 90° bend angle.

Fig. 2.1 shows the stagnation pressure profile as it was at bend angle 0. The lines of constant stagnation pressure were assumed to be horizontal in the theory. In Fig. 2.2 the stagnation pressure profile after 30° bend angle is plotted. According to a linearization the lines of constant stagnation pressure should remain horizontal; they are only approximately so. For the application to the design of

a compressor we are interested in turning angles which are smaller than 30°. However, in order to show how the linearization breaks down the stagnation pressures after 90° bend angle are reproduced in Fig. 2.3. We would have to go still further around the bend in order to be able to recognize the beginning of an oscillating phenomenon.

Our special attention is focussed upon the velocities at 30° bend angle.

The secondary flow which resulted there is plotted on Fig. 3 at $r = \frac{1}{2}(r_i + r_0)$. The full line represents the values obtained by the above theory and the dotted

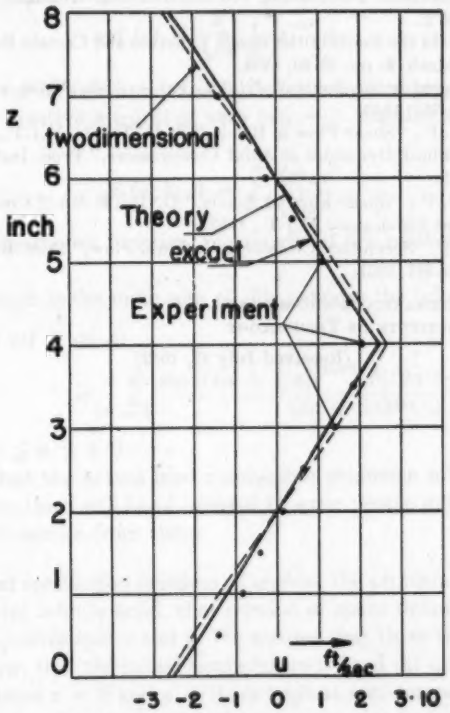


FIG. 3. Velocities in radial direction at 30° bend angle at $r = r_m$ (middle) as a function of the height z.

line is obtained from the simplified theory of Squire and Winter (2) which assumes two-dimensional flow neglecting changes in the direction of increasing bend angle.

It is seen that the experiment which is discussed in more detail in reference (8), agrees with the theory within experimental accuracy. The deviation between the three-dimensional theory and the two-dimensional approximation is of the order of 6 per cent. This means that in the case investigated (a radius to width ratio equal or greater than 28/8) the flow changes so little around the bend

that the simplified theory may be used. For new cases the validity of the simplified theory has to be checked against the three-dimensional theory.

BIBLIOGRAPHY

1. Ainley, D. G., "Analysis of the Air Flow Through Four Stages of Half-Vortex Blading in an Axial Compressor," British Aeronautical Research Council *Reports and Memoranda*, No. 2383 (1950).
2. Squire, H. B. and Winter, K. G., "The Secondary Flow in A Cascade of Airfoils in a Non-Uniform Stream," J. Aeronautical Sciences **18**, pp. 271–277, 1951.
3. Collatz, L., "Numerische Behandlung von Differentialgleichungen," Berlin, Springer, 1951, p. 182 ff.
4. McMahon, S. J., "On the Roots of the Bessel Function and Certain Related Functions," Annals of Math. **9**, pp. 23–30, 1894.
5. Truell, R., "Concerning the Roots of $J_n'(x)N_n'(kx) - J_n'(kx)N_n'(x) = 0$" J. Appl. Phys. **14**, pp. 350–352, 1943.
6. Eichenberger, H. P., "Shear Flow in Bends," Sc.D. Thesis, M.I.T., January 1952.
7. Howell, A. R., "Fluid Dynamics of Axial Compressors," Proc. Inst. Mech. Eng. **153**, p. 441, 1945.
8. Eichenberger, H. P., "Shear Flow in Bends," ONR TR No. 2 Contract N5ori 07848: Gas Turbine Laboratory M.I.T., 1952.
9. Hawthorne, W. R., "Secondary Circulation in Fluid Flow," Proc. Roy. Soc. of London **206**, pp. 374–387, 1951.

Department of Mechanical Engineering
Massachusetts Institute of Technology

(Received July 13, 1952)

A HEAT CONDUCTION PROBLEM AND THE PRODUCT OF TWO ERROR FUNCTIONS

By Earl D. Rainville

1. Introduction. By solving a heat conduction problem in two different ways, and then equating the two solutions, we shall obtain what seems to be a new expansion for the product of two error functions. The error function we use is $\operatorname{erf}(x)$, defined by

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-\alpha^2}\, d\alpha. \tag{1}$$

It will be wise for us to employ the factorial notation, $(a)_n = a(a+1)\cdots(a+n-1)$, for positive integral n, with $(a)_0 = 1$. Another standard notation to be used is

$${}_1F_1(a; b; z) = \sum_{n=0}^{\infty} \frac{(a)_n z^n}{(b)_n n!}, \tag{2}$$

a confluent hypergeometric function, meaningful if b is neither zero nor a negative integer.

The relation which is the main aim of this paper is the following:

$$\operatorname{erf}(r\cos\phi)\operatorname{erf}(r\sin\phi) = \frac{4}{\pi}\sum_{n=0}^{\infty} \frac{\sin[(4n+2)\phi]\, r^{4n+2}\, {}_1F_1(2n+1; 4n+3; -r^2)}{(2n+1)2^{4n+1}(3/2)_{2n}}, \tag{3}$$

valid for $0 < r$, $0 \leqq \phi \leqq \pi/2$.

It is possible that the actual heat conduction problems solved here and the techniques used on them will be of interest to some people independently of the relation (3) which results from them.

2. A simple heat conduction problem. Consider the problem of the conduction of heat in a quarter infinite solid, that portion of space defined by $0 \leqq x$, $0 \leqq y$, in rectangular coordinates x and y. We assume that there is no heat flow normal to the xy-plane, that the initial temperature is fixed (at unity for simplicity) and that the surfaces $x = 0$ and $y = 0$ are kept at zero temperature after some fixed time $t = 0$. Let u represent the temperature at (x, y) at time t.

The physical problem roughly described above can be stated mathematically as the following boundary value problem:

$$\frac{\partial u}{\partial t} = h^2\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right), \qquad \text{for } 0 < t, 0 < x, 0 < y; \tag{4}$$

$$\text{As } t \to 0^+, \quad u \to 1, \qquad \text{for } 0 < x, \quad 0 < y; \tag{5}$$

$$\text{As } x \to 0^+, \quad u \to 0, \qquad \text{for } 0 < t, \quad 0 < y; \tag{6}$$

$$\text{As } y \to 0^+, \quad u \to 0, \qquad \text{for } 0 < t, \quad 0 < x, \tag{7}$$

in which h^2, the thermal diffusivity, is presumed constant.

In an elementary study of the simple heat equation,

$$\partial u/\partial t = h^2\, \partial^2 u/\partial x^2, \tag{8}$$

one of the first solutions encountered is

$$u_1 = \operatorname{erf}[x/(2ht^{\frac{1}{2}})]. \tag{9}$$

The function u_1 has the properties:

(a) It satisfies equation (8) for $0 < x, \quad 0 < t$;

(b) As $t \to 0^+, \quad u_1 \to 1$, for $0 < x$;

(c) As $x \to 0^+, \quad u_1 \to 0$, for $0 < t$.

It follows at once that the solution we seek can be formed as a product of solutions (9), one in x and t, the other in y and t. That is,

$$u = \operatorname{erf}[x/(2ht^{\frac{1}{2}})] \operatorname{erf}[y/(2ht^{\frac{1}{2}})] \tag{10}$$

is a solution of the boundary value problem (4)–(7), a fact easily verified directly. This is not a new result.

3. Solution in polar coordinates. In polar coordinates ρ and θ, related to the x and y of Section 2 by $x = \rho \cos\theta$, $y = \rho \sin\theta$, the boundary value problem (4)–(7) can be restated as follows:

$$\frac{\partial u}{\partial t} = h^2\left(\frac{\partial^2 u}{\partial \rho^2} + \frac{1}{\rho}\frac{\partial u}{\partial \rho} + \frac{1}{\rho^2}\frac{\partial^2 u}{\partial \theta^2}\right), \qquad \text{for } 0 < t, 0 < \theta < \frac{\pi}{2}, 0 < \rho; \tag{11}$$

$$\text{As } t \to 0^+, \qquad u \to 1, \text{ for } 0 < \rho, \qquad 0 < \theta < \pi/2; \tag{12}$$

$$\text{As } \theta \to 0^+, \qquad u \to 0, \qquad \text{for } 0 < t, \qquad 0 < \rho; \tag{13}$$

$$\text{As } \theta \to (\pi/2)^-, \qquad u \to 0, \qquad \text{for } 0 < t, \qquad 0 < \rho. \tag{14}$$

We shall solve the boundary value problem (11)–(14) by first seeking certain types of solutions of the differential equation (11) and then adjusting them to satisfy the boundary conditions (12)–(14). As usual, our technique can be purely formal because we can verify the final result directly in (11)–(14).

For equation (11) we shall obtain solutions which are functions of only two, not three, arguments,

$$u = \psi(-\rho^2/4h^2t, \theta) \equiv \psi(v, \theta). \tag{15}$$

It is possible to look upon this as a change of independent variables in which θ is untouched and new variables v and w are introduced by

$$v = -\rho^2/4h^2t, \text{ and } w = w(\rho, t),$$

with the function w not specified. Then we seek solutions independent of w.

Direct substitution of the u of (15) into equation (11) yields the new equation

$$v^2\frac{\partial^2 \psi}{\partial v^2} + v(1 - v)\frac{\partial \psi}{\partial v} + \frac{1}{4}\frac{\partial^2 \psi}{\partial \theta^2} = 0, \qquad \text{for } v < 0, 0 < \theta < \frac{\pi}{2}. \tag{16}$$

In a similar manner the conditions (12)–(14) become

$$\text{(17)} \qquad \text{As } v \to -\infty, \quad \psi \to 1, \quad \text{for } 0 < \theta < \pi/2;$$

$$\text{(18)} \qquad \text{As } \theta \to 0^+, \quad \psi \to 0, \quad \text{for } v < 0;$$

$$\text{(19)} \qquad \text{As } \theta \to (\pi/2)^-, \quad \psi \to 0, \quad \text{for } v < 0.$$

The boundary value problem (16)–(19) is amenable to the simple technique of separation of variables. Put $\psi = f(v)h(\theta)$. Then

$$\text{(20)} \qquad h'' + \alpha^2 h = 0$$

and

$$\text{(21)} \qquad v^2 f'' + v(1 - v)f' - \tfrac{1}{4}\alpha^2 f = 0.$$

Because of (18) and (19) we find that $\alpha = 2k$, k integral and $h = \sin 2k\theta$. Then f must satisfy the equation

$$v^2 f'' + v(1 - v)f' - k^2 f = 0,$$

the general solution of which is

$$f = Av^k \, {}_1F_1(k; 2k + 1; v) + Bv^{-k} \, {}_1F_1(-k; 1 - 2k; v),$$

with A and B arbitrary. The coefficient of B is a polynomial, of degree k, in $(1/v)$. Since $v \to 0$, as $\rho \to 0$ for fixed t, we must discard such solutions, which leaves us with

$$\text{(22)} \qquad \psi = \sum_{k=1}^{\infty} A_k v^k \, {}_1F_1(k; 2k + 1; v) \sin 2k\theta.$$

The ψ of (22) satisfies (16), (18), and (19) and needs to be adapted to the condition (17). That is, as $v \to -\infty$, the right member should become the Fourier sine series for unity in the interval $0 < \theta < \pi/2$.

It is common knowledge that

$$\text{(23)} \qquad 1 = \frac{4}{\pi} \sum_{n=0}^{\infty} \frac{\sin [(4n + 2)\theta]}{2n + 1}, \qquad \text{for } 0 < \theta < \frac{\pi}{2}.$$

Therefore, in (22) we must choose $A_{2n} = 0$ and require that

$$\text{(24)} \qquad A_{2n+1} \lim_{v \to -\infty} [v^{2n+1} \, {}_1F_1(2n + 1; 4n + 3; v)] = \frac{4}{(2n + 1)\pi}.$$

Asymptotic properties of the ${}_1F_1$ are known. From Copson's Complex Variables, p. 265, we obtain

$$\text{(25)} \qquad \lim_{v \to -\infty} [v^{2n+1} \, {}_1F_1(2n + 1; 4n + 3; v)] = -\frac{\Gamma(4n + 3)}{\Gamma(2n + 2)} = -2^{4n+1}\left(\frac{3}{2}\right)_{2n}.$$

Hence, from (24),

$$A_{2n+1} = \frac{4}{\pi} \frac{-1}{(2n + 1)2^{4n+1}(\frac{3}{2})_{2n}},$$

and our solution (22) is now found to be

$$(26) \qquad \psi = \frac{4}{\pi}\sum_{n=0}^{\infty} \frac{(-v)^{2n+1}\,{}_1F_1(2n+1;\,4n+3;\,v)\,\sin\,[(4n+2)\theta]}{(2n+1)2^{4n+1}(\tfrac{3}{2})_{2n}}.$$

That (26) does satisfy the boundary value problem (16)–(19) is easily verified directly. Recalling that $v = -\rho^2/(4h^2t)$, we find that we have the solution

$$(27) \qquad u = \frac{4}{\pi}\sum_{n=0}^{\infty} \frac{[\rho^2/(4h^2t)]^{2n+1}\,{}_1F_1(2n+1;\,4n+3;\,-\rho^2/(4h^2t))\,\sin\,[(4n+2)\theta]}{(2n+1)2^{4n+1}(\tfrac{3}{2})_{2n}},$$

for the original problem (11)–(14).

Before equating the two solutions (27) and (10) of the same problem, it pays to simplify matters by placing

$$x = 2hrt^{\frac{1}{2}}\cos\phi, \qquad y = 2hrt^{\frac{1}{2}}\sin\phi,$$

from which

$$\rho^2 = x^2 + y^2 = 4h^2r^2t$$

and $\theta = \phi$. Then (10) becomes

$$u = \operatorname{erf}(r\cos\phi)\operatorname{erf}(r\sin\phi)$$

and (27) becomes

$$u = \frac{4}{\pi}\sum_{n=0}^{\infty} \frac{r^{4n+2}\,{}_1F_1(2n+1;\,4n+3;\,-r^2)\,\sin\,[(4n+2)\phi]}{(2n+1)2^{4n+1}(\tfrac{3}{2})_{2n}}$$

and the desired relation (3), given earlier, is obtained.

4. An independent verification of the identity (3). A demonstration of the validity of the identity (3) would normally take the form of a uniqueness theorem for the boundary value problem, which was solved in two ways above. We can, however, circumvent the need for such a theorem in a reasonably elementary manner. The function

$$\operatorname{erf}(r\cos\phi)\operatorname{erf}(r\sin\phi)$$

has the Fourier series expansion

$$(28) \qquad \operatorname{erf}(r\cos\phi)\operatorname{erf}(r\sin\phi) = \sum_{n=0}^{\infty}\lambda_n(r)\sin(4n+2)\phi.$$

valid in $0 \leqq \phi \leqq \pi/2$, for all finite r, where

$$(29) \qquad \lambda_n(r) = \frac{4}{\pi}\int_0^{\pi/2} \operatorname{erf}(r\cos\phi)\operatorname{erf}(r\sin\phi)\sin(4n+2)\phi\,d\phi.$$

In the expansion (28) the absence of terms involving $\sin(4n\,\phi)$ is due to the fact that the function being represented is symmetric about $\pi/4$ in the basic interval $0 \leqq \phi \leqq \pi/2$.

The integral in (29) can be evaluated by methods often taught in advanced

calculus, combined with a slight knowledge of the ${}_1F_1$ function. The result is, naturally, that stated in equation (3).

5. A less restricted heat problem. The solution using polar coordinates, of the heat problem of Section 2 is easily extended to handle the situation of a two dimensional wedge of arbitrary angle β. Let the wedge have an initial temperature unity throughout and let its faces $\theta = 0$, $\theta = \beta$, be held at temperature zero for $t > 0$.

The corresponding boundary value problem,

$$\frac{\partial u}{\partial t} = k^2\left(\frac{\partial^2 u}{\partial \rho^2} + \frac{1}{\rho}\frac{\partial u}{\partial \rho} + \frac{1}{\rho^2}\frac{\partial^2 u}{\partial \theta^2}\right), \qquad \text{for } 0 < t, 0 < \rho, 0 < \theta < \beta; \tag{30}$$

$$\text{As } t \to 0^+, \quad u \to 1, \quad \text{for } 0 < \rho, \quad 0 < \theta < \beta; \tag{31}$$

$$\text{As } \theta \to 0^+, \quad u \to 0, \quad \text{for } 0 < t, \quad 0 < \rho; \tag{32}$$

$$\text{As } \theta \to \beta^-, \quad u \to 0, \quad \text{for } 0 < t, \quad 0 < \rho, \tag{33}$$

can be solved by the device of Section 2. In that way it is found that the problem (30)–(33) has the solution

$$u = \frac{4}{\pi}\sum_{n=0}^{\infty} a_n G_n(\rho) \sin\left[(2n+1)\pi\theta/\beta\right], \tag{34}$$

in which

$$G_n(\rho) = \left[\frac{\rho^2}{4k^2 t}\right]^{(2n+1)\pi/\beta} \cdot {}_1F_1\left(\frac{(2n+1)\pi}{2\beta}; 1 + \frac{(2n+1)\pi}{\beta}; \frac{-\rho^2}{4k^2 t}\right),$$

and

$$a_n = \frac{(2n+1)\Gamma\left(1 + \frac{(2n+1)\pi}{2\beta}\right)}{\Gamma\left(1 + \frac{(2n+1)\pi}{\beta}\right)}.$$

THE UNIVERSITY OF MICHIGAN

(Received May 7, 1952)

ON A CLASS OF STOCHASTIC OPERATORS*

By L. A. Zadeh

1. Introduction. By a *stochastic operator* we mean a random function $H(f)$ defined on a function space $\Omega = \{f(t)\}$, where t represents the time. From the physical point of view, a stochastic operator corresponds essentially to a randomly varying system, with the restriction that the operational characteristics of the system are not deterministically related to its operand.

In this paper we are concerned with linear stochastic operators which admit a representation of the form

$$H(f) = \int_{-\infty}^{\infty} H(t, \lambda)\, e^{i\Omega_\lambda}\, dF(\lambda) \tag{1}$$

where $\{H(t, \lambda), -\infty < t < \infty, -\infty < \lambda < \infty\}$ is a complex-valued stationary (wide sense) and continuous (in mean square) stochastic process; λ is a real-valued parameter; $H(t, \lambda)$ is a real function of λ; $F(\lambda)$ is the Fourier-Stieltjes transform of the operand, $f = f(t)$; and the integral is taken in the usual stochastic (Lebesgue) sense.[1] The assumptions on the $\{H(t, \lambda)\}$ process imply that its autocorrelation function $R_H(\tau, \lambda)$,

$$R_H(\tau, \lambda) = E\{H(t, \lambda)\, \overline{H(t + \tau, \lambda)}\} \tag{2}$$

where E denotes the expectation operator and the bar indicates the complex conjugate, is (a) independent of t, (b) continuous at $\tau = 0$, and (c) an even function of τ and λ. In the sequel the autocorrelation function $R_H(\tau, \lambda)$ plays a central role and is referred to as the autocorrelation function of the stochastic operator H.

For each fixed λ, the $\{H(t, \lambda)\}$ process is a continuous stationary (wide sense) stochastic process. As such it may be represented[2] as a continuous path in a separable Hilbert space $L_2(H)$ which consists of all random variables of the form

$$x = \sum_{\nu=1}^{n} c_\nu H(t_\nu, \lambda) \tag{3}$$

where the c's are complex constants and n is any positive integer, and of random variables which are limits (in mean square) of sequences of x's. The inner product in this space is defined as

$$(x, y) = E\{x\bar{y}\}. \tag{4}$$

Accordingly, the autocorrelation function $R_H(\tau, \lambda)$ may be regarded as the inner product of $H(t, \lambda)$ and $H(t + \tau, \lambda)$ in $L_2(H)$.

* Presented to the American Mathematical Society, April 27, 1951, under the title *On Stochastic Heaviside Operators*.

[1] P. Levy, *Processus Stochastiques et Mouvement Brownien*, Gauthier-Villars, Paris, 1948.

[2] K. Karhunen, *Über Lineare Methoden in der Wahrscheinlichkeitsrechnung*, Ann. Ac. Sci. Fennicae, A I, **37**, 1947.

It will be noted that when $H(t, \lambda)$ is a polynomial in λ, H reduces to a stochastic differential operator. It will also be noted that the random function $H(t, \lambda)$ may be expressed in terms of H by the relation

$$H(t, \lambda) = e^{-i\Omega} H(e^{i\Omega}). \tag{5}$$

2. A relation between autocorrelation functions. In this section we consider the case where $\{f(t), -\infty < t < \infty\}$ is a real-valued stationary (wide sense) and continuous stochastic process with a given autocorrelation function, $R_f(\tau)$. We assume further that the $\{f(t)\}$ and $\{H(t, \lambda)\}$ processes are independent. Proceeding on these assumptions, we derive an expression for the autocorrelation function, $R_g(\tau)$, of the transform of f,

$$g = H(f), \tag{6}$$

and show that R_g may be written as

$$R_g = R_H(R_f) \tag{7}$$

where the transformation is defined as in (1), with R_H, $R_H(\tau, \lambda)$ and $R_f(\tau)$ playing the roles of H, $H(t, \lambda)$ and $f(t)$ respectively. This result was stated without proof in a previous paper.[3]

In the sequel, we shall find it necessary on occasion to express the operation with H on f in a form that brings in evidence the function $H(t, \lambda)$. In such cases, we shall write (6) as

$$g(t) = H(t, \theta)f(t) \tag{8}$$

where the use of θ in place of λ indicates that $H(t, \theta)$ should be regarded as an operator acting on $f(t)$ in accordance with the defining equation (1). In short,

$$H(f) = H(t, \theta)f(t) = \int_{-\infty}^{\infty} H(t, \lambda)e^{i\Omega}\, dF(\lambda) \tag{9}$$

where $F(\lambda)$ is the Fourier-Stieltjes transform of $f(t)$.

Turning to the autocorrelation functions of $f(t)$, $H(t, \lambda)$ and $g(t)$, we have

$$R_f(\tau) = E\{f(t)f(t + \tau)\} \tag{10}$$

$$R_H(\tau, \lambda) = E\{H(t, \lambda)\, \overline{H(t + \tau, \lambda)}\} \tag{11}$$

and

$$R_g(\tau) = E\{g(t)g(t + \tau)\}. \tag{12}$$

Note that $\{g(t)\}$ is a real-valued stochastic process by virtue of the assumption that $H(t, \lambda)$ is a real-valued function of $i\lambda$.

Proceeding in a formal manner, we have

$$g(t)g(t + \tau) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} H(t, \lambda)H(t + \tau, \lambda')e^{i\Omega} e^{i(t+\tau)\lambda'}\, dF(\lambda)\, dF(\lambda'). \tag{13}$$

[3] L. A. Zadeh, *Time-Dependent Heaviside Operators*, Journ. Math. Phys., **30**, 73, 1951.

On applying the expectation operator to both sides of (13) and making use of the commutativity of stochastic averaging and stochastic integration, we obtain

$$R_g(\tau) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} E\{H(t, \lambda)H(t + \tau, \lambda')\}e^{i\Omega}e^{i((t+\tau)\lambda'} E\{dF(\lambda)\, dF(\lambda')\}. \quad (14)$$

Since $\{f(t)\}$ is a continuous stationary process, the $\{F(\lambda)\}$ process associated with its Fourier-Stieltjes representation

$$f(t) = \int_{-\infty}^{\infty} e^{i\Omega}\, dF(\lambda) \quad (15)$$

is a process with uncorrelated increments.[4] This implies that

$$E\{dF(\lambda)\, dF(\lambda')\} = 0 \qquad \text{for } \lambda \neq -\lambda' \quad (16)$$

and

$$E\{|\, dF(\lambda)\,|^2\} = d\sigma_f(\lambda) \quad (17)$$

where $\sigma_f(\lambda)$ is the spectral distribution function of the $\{f(t)\}$ process.

In consequence of (16) and (17), (14) reduces to

$$R_g(\tau) = \int_{-\infty}^{\infty} E\{H(t, \lambda)H(t + \tau, -\lambda)\}e^{-i\tau\lambda}\, d\sigma_f(\lambda), \quad (18)$$

and since $H(t, i\lambda)$ is a real-valued function, we have

$$H(t + \tau, -\lambda) = \overline{H(t + \tau, \lambda)}, \quad (19)$$

and hence

$$E\{H(t, \lambda)H(t + \tau, -\lambda)\} = R_H(\tau, \lambda) \quad (20)$$

where $R_H(\tau, \lambda)$ is the autocorrelation function of H. Finally, on substituting (20) in (18) and replacing τ by $-\tau$, we obtain

$$R_g(\tau) = \int_{-\infty}^{\infty} R_H(\tau, \lambda)e^{i\tau\lambda}\, d\sigma_f(\lambda) \quad (21)$$

where $\sigma_f(\lambda)$ is the Fourier-Stieltjes transform of $R_f(\tau)$.

We note that this relation is of the same form as (1), with $R_H(\tau, \lambda)$ and $\sigma_f(\lambda)$ corresponding to $H(t, \lambda)$ and $F(\lambda)$, respectively. The role of the operand in (21) is played by $R_f(\tau)$, and that of the operator by R_H.

On the basis of this similarity, we may write (21) in the operational form (8)

$$R_g(\tau) = R_H(\tau, \theta)R_f(\tau) \quad (22)$$

with the understanding that (22) is defined by (21).

To summarize: If

$$g(t) = H(t, \theta)f(t) \quad (23)$$

where $\{f(t), -\infty < t < \infty\}$ is a real-valued continuous stationary (wide sense) process; $H(t, \lambda)$ is a real function of $i\lambda$; $\{H(t, \lambda), -\infty < t < \infty, -\infty < \lambda < \infty\}$

[4] J. L. Doob, *Time Series and Harmonic Analysis*, Proc. Berkeley Symposium on Math. Stat. and Probability, 303, 1949.

is a continuous stationary (wide sense) process independent of $\{f(t)\}$; and (23) is defined by (9); then

$$R_g(\tau) = R_H(\tau, \theta)R_f(\tau) \tag{24}$$

where R_g, R_H and R_f are the autocorrelation functions of $\{g(t)\}$, $\{H(t, \lambda)\}$ and $\{f(t)\}$ respectively, and (24) is defined by (21), which is of the same form as (1).

It should be remarked that (24) is also valid when the autocorrelation functions R_g, R_H and R_f are defined as the temporal rather than stochastic means of $g(t)g(t + \tau)$, $H(t, \lambda)\overline{H(t + \tau, \lambda)}$ and $f(t)f(t + \tau)$ respectively. This makes a difference when the $\{f(t)\}$ and $\{H(t, \lambda)\}$ processes are not ergodic. In this case, we require that the temporal means of $f(t)f(t + \tau)$ and $H(t, \lambda)\overline{H(t + \tau, \lambda)}$ be independent of t. The requirement that the $\{f(t)\}$ and $\{H(t, \lambda)\}$ processes be continuous may be dropped if one will admit the use of delta-functions of various orders in conjunction with (24).

As a simple application of the result derived above, we consider the following process

$$g(t) = f(t + \alpha(t)) \tag{25}$$

where $\{\alpha(t), -\infty < t < \infty\}$ and $\{f(t), -\infty < t < \infty\}$ are continuous stationary (wide sense) and independent stochastic processes.

It is readily verified that (25) may be written in the operational form

$$g(t) = H(t, \theta)f(t) \tag{26}$$

where

$$H(t, \theta) = e^{i\alpha(t)\theta}. \tag{27}$$

By (2), the autocorrelation function of the stochastic operator associated with $H(t, \theta)$ is given by

$$R_H(\tau, \lambda) = E\{\exp [i\alpha(t)\lambda - i\alpha(t + \tau)\lambda]\}, \tag{28}$$

which in terms of the characteristic function of the joint distribution of $\alpha(t)$ and $\alpha(t + \tau)$,

$$M_\alpha(\lambda, \lambda' \tau;) = E\{\exp [i\alpha(t)\lambda + i\alpha(t + \tau)\lambda']\} \tag{29}$$

may be written as

$$R_H(\tau, \lambda) = M_\alpha(\lambda, -\lambda; \tau). \tag{30}$$

On substituting this in (24), the expression for $R_g(\tau)$ becomes

$$R_g(\tau) = M_\alpha(\theta, -\theta; \tau)R_f(\tau). \tag{31}$$

In particular, if $\{\alpha(t)\}$ is a Gaussian process with zero mean, the characteristic function is given by[5]

$$M_\alpha(\lambda, \lambda'; \tau) = \exp \{-R_\alpha(0)\tfrac{1}{2}(\lambda^2 + \lambda'^2) - R_\alpha(\tau)\lambda\lambda'\} \tag{32}$$

[5] S. O. Rice, *Mathematical Analysis of Random Noise*, Bell Sys. Tech. Journ., **23**, 282, 1944.

where $R_\alpha(\tau)$ is the autocorrelation function of the $\{\alpha(t)\}$ process. From (30) it follows at once that

$$R_H(\tau, \lambda) = \exp \{\lambda^2[R_\alpha(\tau) - R_\alpha(0)]\} \tag{33}$$

and hence the autocorrelation function of the $\{g(t)\}$ process, (25), with Gaussian $\{\alpha(t)\}$, is given by

$$R_g(\tau) = \exp \{\theta^2[R_\alpha(\tau) - R_\alpha(0)]\} R_f(\tau) \tag{34}$$

or more explicitly

$$R_g(\tau) = \int_{-\infty}^{\infty} \exp \{\lambda^2[R_\alpha(\tau) - R_\alpha(0)]\} e^{i\tau\lambda} \, d\sigma_f(\lambda) \tag{35}$$

where $\sigma_f(\lambda)$ is the spectral distribution function of the $\{f(t)\}$ process.

3. A product relation for autocorrelation functions of stochastic operators. If H_1 and H_2 are independent stochastic operators of the type defined by (1), then their product, $H = H_1H_2$, is a stochastic operator H defined by the equation $H(f) = H_1[H_2(f)]$. The domain of H is the set of the elements of the domain of H_2 which are transformed by H_2 into elements of the domain of H_1.

Let $H(t, \lambda)$, $H_1(t, \lambda)$ and $H_2(t, \lambda)$ be the random functions associated with H, H_1 and H_2 respectively. By (5), we have

$$H(t, \lambda) = e^{-i\Omega} H(e^{i\Omega}) = e^{-i\Omega} H_1[H_2(e^{i\Omega})]. \tag{36}$$

But

$$H_2(e^{i\Omega}) = e^{i\Omega} H_2(t, \lambda) \tag{37}$$

and hence

$$H(t, \lambda) = e^{-i\Omega} H_1[e^{i\Omega} H_2(t, \lambda)] \tag{38}$$

or more explicitly

$$H(t, \lambda) = e^{-i\Omega} H_1(t, \theta)[e^{i\Omega} H_2(t, \lambda)]. \tag{39}$$

Now it is readily verified that

$$H_1(t, \theta) [e^{i\Omega} H_2(t, \lambda)] = e^{i\Omega} H_1(t, \theta + \lambda) H_2(t, \lambda), \tag{40}$$

and consequently the random function $H(t, \lambda)$ is given by

$$H(t, \lambda) = H_1(t, \theta + \lambda)H_2(t, \lambda). \tag{41}$$

In this relation, λ is a fixed parameter and $H_2(t, \lambda)$ plays the role of the operand of $H_1(t, \theta + \lambda)$.

Let $R_H(\tau, \lambda)$, $R_{H_1}(\tau, \lambda)$ and $R_{H_2}(\tau, \lambda)$ be the autocorrelation functions of H, H_1 and H_2 respectively. On applying (24) to (41) and treating $H(t, \lambda)$, $H_1(t, \theta + \lambda)$ and $H_2(t, \lambda)$ as $g(t)$, $H(t, \theta)$ and $f(t)$, respectively, we obtain the following relation between the autocorrelation functions in question

$$R_H(\tau, \lambda) = R_{H_1}(\tau, \theta + \lambda)R_{H_2}(\tau, \lambda) \tag{42}$$

where λ is a fixed parameter and $R_{H_2}(\tau, \lambda)$ is the operand of $R_{H_1}(\tau, \theta + \lambda)$.

Since (42) is of the same form as (41), we may conclude that R_H, regarded as an ordinary operator defined by (21), is given by the product of R_{H_1} and R_{H_2}

$$R_H = R_{H_1} R_{H_2} \tag{43}$$

where R_{H_1} and R_{H_2} are likewise regarded as ordinary operators defined by (21).

To summarize: If H is the product of two independent stochastic operators H_1 and H_2, $H = H_1 H_2$, then

$$H(t, \lambda) = H_1(t, \theta + \lambda) H_2(t, \lambda) \tag{44}$$

and

$$R_H(\tau, \lambda) = R_{H_1}(\tau, \theta + \lambda) R_{H_2}(\tau, \lambda) \tag{45}$$

or symbolically, $R_H = R_{H_1} R_{H_2}$, where R_H, R_{H_1} and R_{H_2} are the autocorrelation functions of H, H_1, and H_2, respectively.

This result furnishes an effective means of determining the autocorrelation function of a stationary (wide sense) stochastic process which admits a representation of the form

$$g = H_1 H_2 \cdots H_n(f) \tag{46}$$

where $H_1, \cdots, H_n$ and f are mutually independent.

COLUMBIA UNIVERSITY

(Received April 11, 1951)

UPPER AND LOWER BOUNDS FOR TORSIONAL RIGIDITY*

By H. F. Weinberger

1. Introduction. We are here concerned with the Saint-Venant [7] torsion problem of a long cylindrical beam of uniform cross section.

The axis of the beam is taken as the z-axis. The cross section is a bounded region R in the xy plane. Let this region be bounded by an outer curve γ_0 and possibly one or more inner curves γ_i (holes).

The state of stress in the interior of the beam is determined by the stress function $\Psi(x, y)$ defined on R. The stress is given by

$$\tau_{zx} = \mu\alpha \frac{\partial \Psi}{\partial y}, \qquad \tau_{zy} = -\mu\alpha \frac{\partial \Psi}{\partial x} \tag{1}$$

where μ is the shear modulus and α is the angle of twist per unit length.

The function Ψ is defined by

$$\Delta\Psi = -2 \qquad \Psi = 0 \text{ on } \gamma_0, \qquad \Psi = C_i \text{ on } \gamma_i \quad i = 1, \cdots, n. \tag{2}$$

The constants C_i, which occur only when R is multiply connected, are to be determined by the additional condition

$$\oint_\gamma \frac{\partial \Psi}{\partial n}\, ds = -2A(\gamma) \tag{3}$$

for any closed curve γ in R, where $A(\gamma)$ is the area inclosed by the curve γ, and n is the outward normal on γ. If R is simply connected, (3) is a consequence of Green's theorem. If R is multiply connected, Green's theorem assures us that (3) is satisfied if it holds for each of the interior boundary curves.

The torsional rigidity S, which is the torque required for unit angle of twist per unit length when $\mu = 1$, is the Dirichlet integral of Ψ.

$$S = \iint_R (\Psi_x^2 + \Psi_y^2)\, dx\, dy. \tag{4}$$

The explicit solution of the boundary value problem (2), (3) often presents great difficulties. It is therefore desirable to approximate S even when Ψ is not known.

A method for obtaining upper and lower bounds for S with arbitrary accuracy has been given by Diaz and Weinstein [4, 5], and in a more complete form by Diaz [2, 3], as an extension of the methods of Rayleigh-Ritz and Trefftz.

This paper presents new methods for finding upper and lower bounds for a restricted class of cross sections. For these problems, the bounds are obtained with great ease and speed. While their accuracy cannot be systematically improved, it is often quite good, particularly for regions with small holes. For a hollow square with sufficiently small hole we obtain better bounds than those which have so far been calculated by Diaz [3].

* This work was sponsored by the Office of Naval Research.

2. A minimum principle. We shall use the following extension to multiply connected domains of a minimum principle given by Diaz [2]:

$$S = \min \iint_R (f_x^2 + f_y^2)\, dx\, dy \tag{5}$$

among continuously differentiable functions $f(x, y)$ satisfying $\Delta f = 0$ and

$$\oint_\gamma \frac{\partial f}{\partial n}\, ds = -2A(\gamma)$$

for all closed curves γ in R.

To prove this, we let $\Phi = f - \Psi$. Then

$$\Delta\Phi = 0, \qquad \oint_\gamma \frac{\partial \Phi}{\partial n}\, ds = 0 \text{ for all } \gamma \text{ in } R. \tag{6}$$

Hence,

$$\begin{aligned}
&\iint_R (f_x^2 + f_y^2)\, dx\, dy \\
&\quad = \iint_R (\Psi_x^2 + \Psi_y^2)\, dx\, dy + \iint_R (\Phi_x^2 + \Phi_y^2)\, dx\, dy \\
&\qquad + 2 \iint_R (\Phi_x \Psi_x + \Phi_y \Psi_y)\, dx\, dy \\
&\quad = S + \iint_R (\Phi_x^2 + \Phi_y^2)\, dx\, dy - 2 \sum_{i=1}^{n} \oint_{\gamma_i} \Psi \frac{\partial \Phi}{\partial n}\, ds \\
&\quad = S + \iint_R (\Phi_x^2 + \Phi_y^2)\, dx\, dy - 2 \sum_{i=1}^{n} C_i \oint_{\gamma_i} \frac{\partial \Phi}{\partial n}\, ds \\
&\quad = S + \iint_R (\Phi_x^2 + \Phi_y^2)\, dx\, dy
\end{aligned} \tag{7}$$

from which (5) follows.

3. An inequality. We consider a region R_{12} which is the union of two disjoint regions R_1 and R_2. Let their respective torsional rigidities be S_{12}, S_1, and S_2. We show that

$$S_{12} \geqq S_1 + S_2. \tag{8}$$

Let Ψ be the stress function for R_{12}. Then it satisfies $\Delta\Psi = -2$,

$$\oint_\gamma \frac{\partial \Psi}{\partial n}\, ds = -2A(\gamma)$$

for all γ in R_{12}; hence also in R_1 and in R_2. Therefore, by (4) and (5)

$$\begin{aligned}
S_{12} &= \iint_{R_{12}} (\Psi_x^2 + \Psi_y^2)\, dx\, dy \\
&= \iint_{R_1} (\Psi_x^2 + \Psi_y^2)\, dx\, dy + \iint_{R_2} (\Psi_x^2 + \Psi_y^2)\, dx\, dy \\
&\geqq S_1 + S_2.
\end{aligned} \tag{9}$$

From (7) it follows that the equality holds if and only if Ψ only differs by constants from the stress functions in R_1 and in R_2, that is, if Ψ is constant on the common boundary of R_1 and R_2. In other words, *the equality in* (8) *holds if and only if the common boundary of* R_1 *and* R_2 *is a level line of* Ψ.

Obviously, we can also show for the union of subregions

$$S_{1\ldots n} \geqq S_1 + \cdots + S_n . \tag{10}$$

This permits us to obtain a lower bound for $S_{1\ldots n}$. For example, lower bounds for L, T, or I shaped beams are easily obtained by splitting them into rectangles. The numerical results are not always satisfactory, and an alternative lower bound for certain problems will be given in paragraph 5.

Since the torsional rigidity is always positive, (10) will still hold if some of the terms on the right are replaced by zero.

4. Upper bounds. Inequality (10) can also be used to get upper bounds for the torsional rigidity S_1. Transposing (10) gives

$$S_1 \leqq S_{1\ldots n} - (S_2 + \cdots + S_n) \tag{11}$$

This formula is particularly useful for a multiply connected region R_1 consisting of a simply connected region $R_{1\ldots n}$ with holes $R_2, \ldots, R_n$, all having known rigidities.

As an example, consider a hollow square with outer side $2b$ and inner side $2a$. The torsional rigidity of the solid square is $2.2496b^4$ and that of the hole is $2.2496a^4$. Thus, we have for the hollow square

$$S \leqq 2.2496(b^4 - a^4) \tag{12}$$

Since the equality holds for $a = 0$, (12) is the best upper bound of the form $C(b^4 - a^4)$. For $a < b/4$, (12) gives a better bound than that given by Diaz [3] and Weber [8].

We note that (12) holds even if the square hole is not concentric with or parallel to the outer square.

For the problem proposed by Courant [1] of a square with four square holes we immediately find

$$S \leqq 2.2496(b^4 - 4\,a^4). \tag{13}$$

5. Lower bounds. In this paragraph we give a second method for obtaining lower bounds, which gives better results than (8) for certain multiply connected regions. The method depends on the following theorem.

THEOREM 1. Let R_1 consist of the simply connected region R_{12} with the hole R_2. Let Ψ be the stress function of R_{12}. Let R' be a region contained in R_{12} and containing R_2 such that there is a level line of Ψ lying entirely outside R_2 but inside R_2'. We call R_2'' the region inside this level line and R_1'' the region outside it but still in R_{12}. The corresponding torsional rigidities of the various regions are labelled in the obvious manner. Then

$$S_1 \geqq S_{12} - S_2' . \tag{14}$$

We return to the statement in paragraph 3 to the effect that if R_{12} is cut along a level line of Ψ, the equality holds in (8). This immediately gives

$$S_1'' = S_{12} - S_2''. \tag{15}$$

Since R_1 contains R_1'',

$$S_1 \geqq S_1''. \tag{16}$$

Since R_2' contains R_2'',

$$S_2' \geqq S_2''. \tag{17}$$

Combining (15), (16), and (17) gives (14) and the theorem is proved.

To apply this theorem, one must find a region R_2' having the properties described in the theorem.

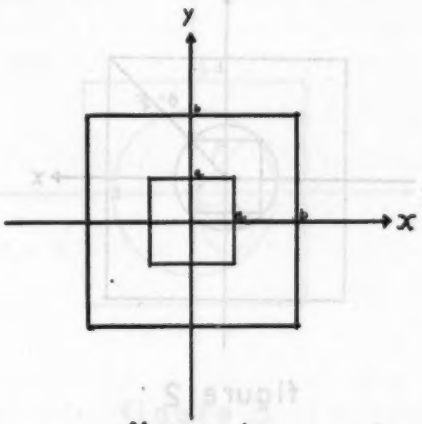


figure 1

We again consider the hollow square with sides $2b$ and $2a$, and assume that $a < b/\sqrt{2}$. We show by means of the following two lemmas that the circle circumscribed about the hole may be used as R_2'.

LEMMA 1. Let Ψ be stress function of the (solid) square of side $2b$. Of all the values of Ψ attained on the smaller square, the minimum occurs at the corners.

Proof. Since $\Delta\Psi = -2$, $\Delta(\Psi_x) = 0$.

Since Ψ is symmetric about the axes, $\Psi_x = 0$ at $x = 0$. Because of the boundary conditions, $\Psi_x = 0$ at $y = \pm b$. Finally, since $\Delta\Psi < 0$, Ψ cannot have a minimum in the interior. But $\Psi = 0$ on the boundary and hence must be positive inside. So $\Psi_x \leqq 0$ at $x = b$. Thus, $\Psi_x \leqq 0$ on the boundary of the rectangle $0 \leqq x \leqq b$, $-b \leqq y \leqq b$ and is harmonic in its interior. So $\Psi_x < 0$ inside this rectangle. Thus, Ψ steadily decreases from its maximum at $x = 0$ along the lines $x = \pm a$. By symmetry, the same is true on the lines $y = \pm a$, and the statement of the lemma follows.

LEMMA 2. Of all the values of Ψ attained on the circle inscribed about the inner square, the maximum occurs at the corners of the square.

Proof. Introduce polar coordinates (r, θ). Since $\Delta\Psi = -2$, $\Delta(\Psi_\theta) = 0$.

Since Ψ is symmetric about the y-axis as well as about the line $\theta = \pi/4$, $\Psi_\theta = 0$ on each of these lines. Since the direction of increasing θ is outward on $y = b$, $x \geqq 0$, and since Ψ is positive inside, we find that $\Psi_\theta \leqq 0$ on this part of the boundary. Thus, in the triangle $0 < y < b$, $\pi/4 < \theta < \pi/2$, $\Psi_\theta < 0$. Hence, Ψ steadily decreases in value along the circle from the diagonal to the vertical position. Together with symmetry, this proves the lemma.

Consider the level line of Ψ passing through the corners of the inner square. By the two lemmas, the value of Ψ on this level line remains between the values of Ψ on the inner square and those on the circumscribed circle. Hence the level

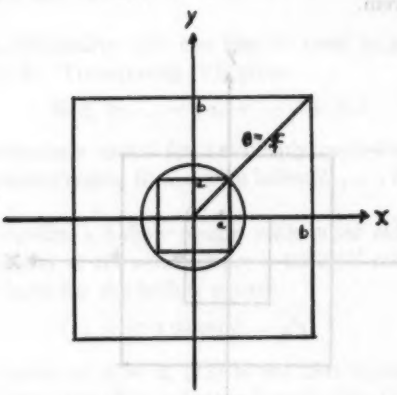


figure 2

line remains between these two curves. We have thus shown that the circumscribed circle, which has radius $a\sqrt{2}$, is a region R_2' whose rigidity S_2' can be used in the inequality (17). By this means, we get for the hollow square the lower bound

$$S \geqq 2.2496b^4 - 2\pi a^4. \tag{18}$$

For $a < b/4$ this bound is better than the best lower bound in the literature [8]. Furthermore, the difference between this lower bound and the upper bound (12) is $4.0336a^4$, which is quite small relative to the bounds for small a/b.

It is immediately obvious from the proofs of the two lemmas that for the problem of the square with a concentric circular hole we may take for R_2' the square parallel to the outer square circumscribed about the hole. Thus, for a square of side $2b$ with hole of radius r, we find

$$S \geqq 2.2496(b^4 - r^4). \tag{19}$$

Furthermore, lemmas 1 and 2 can clearly be extended to any regular polygon. Thus, we find lower bounds for regular polygons with similar holes or circular holes.

If a given doubly connected region is not of the type for which we have here obtained lower bounds, we may deform it by enlarging the hole or shrinking the outer boundary. Then the lower bound for the deformed region is a bound for the original region.

If the stress function Ψ is explicitly known, R_1' may, of course, be taken as just the region inside any level line remaining outside the hole.

Lower bounds may sometimes be found by combining the technique of this paragraph with the inequality (8). Consider the problem of a square with side $2b$ having a circular hole of radius a with center at distance e from the center

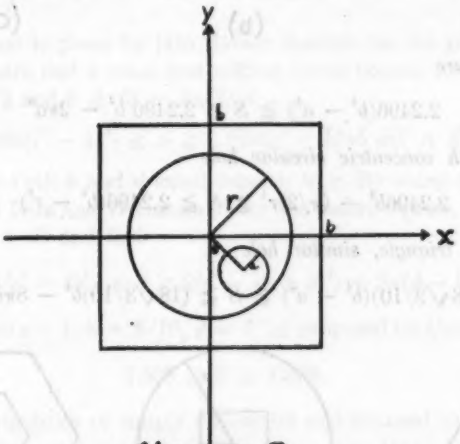


figure 3

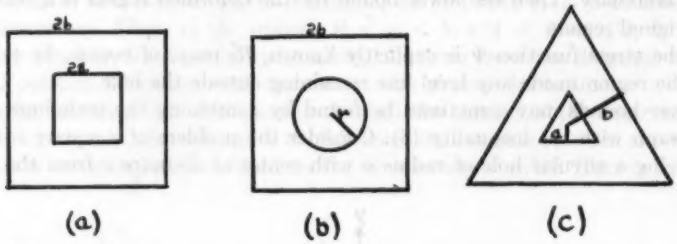
of the square. We introduce a cut consisting of a circle of radius r concentric with the square and surrounding the hole. It divides our region into two subregions. One of these is the square with concentric circular hole, for which we have the lower bound (19). The other is a circle with eccentric circular hole, whose torsional rigidity is known [9]. According to (8), the sum of the rigidities of the two regions is a lower bound for the rigidity of the original region. This lower bound still, of course, depends on the choice of the radius r. As an example, when we take $b = 1$, $a = 0.2$, $e = 0.1$, the best result is obtained with $r = 0.3$ (its smallest possible value). We find for this case

$$S \geqq 2.23. \tag{20}$$

The upper bound (11) gives

$$S \leqq 2.25. \tag{21}$$

6. Table of results. For convenience, we have tabulated upper and lower bounds which have been calculated by the methods treated in this paper. It is to be noted that in each case the two bounds become equal as the hole decreases in size to zero.



(a) (b) (c)

(a) *Hollow square*

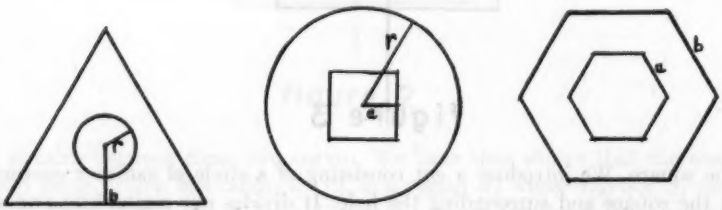
$$2.2496(b^4 - a^4) \geqq S \geqq 2.2496\, b^4 - 2\pi a^4$$

(b) *Square with concentric circular hole*

$$2.2496b^4 - (\pi/2)r^4 \geqq S \geqq 2.2496(b^4 - r^4)$$

(c) *Equilateral triangle, similar hole*

$$(18\sqrt{3}/10)(b^4 - a^4) \geqq S \geqq (18\sqrt{3}/10)b^4 - 8\pi a^4$$



(d) (e) (f)

(d) *Equilateral triangle, circular hole*

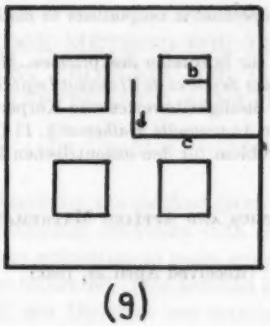
$$(18\sqrt{3}/10)b^4 - (\pi/2)r^4 \geqq S \geqq (18\sqrt{3}/10)(b^4 - 16r^4)$$

(e) *Circle, square hole*

$$(\pi/2)r^4 - 2.2496a^4 \geqq S \geqq (\pi/2)r - 2\pi a^4$$

(f) *Regular hexagon, similar hole*

$$1.0359(b^4 - a^4) \geqq S \geqq 1.0359b^4 - (\pi/2)a^4$$



(g) *Square with four square holes.*

The upper bound is given by (13). Lower bounds can be given by splitting into a hollow square and a cross and adding lower bounds for their rigidities. If $b + d \leqq c/\sqrt{2}$ and $b \geqq (3 - 2\sqrt{2})\, d$,

$$2.2496(c^4 - 4b^4) \geqq S \geqq 2.2496c^4 - 16\pi b\, d(b^2 + d^2),$$

which is good for both b and d small relative to c. By using the lower bound $(c^4 - (b + d)^4)$ of Diaz and Weinstein [5] for the hollow square, we can drop the restriction on $(b + d)$ and find

$$2.2496(c^4 - 4b^4) \geqq S \geqq 2(c^4 - (b + d)^4) + 2\pi(d - b)^4$$

For the dimensions $c = 1$, $b = 5/16$, $d = 7/16$ proposed by Courant [1], we find

$$1.538 \geqq S \geqq 1.369.$$

The torsional rigidities of simply connected regions used in finding the numerical bounds above and throughout this paper were taken from the tables in Polya and Szegö [6].

The author wishes to thank L. E. Payne and E. H. Zarantonello for a great deal of aid and discussion on the topics here presented.

BIBLIOGRAPHY

1. Courant, R. Variational Methods for the Solution of Problems of Equilibrium and Vibrations. *Bulletin of the American Math. Soc.* 49 (1943) p. 1.
2. Diaz, J. B. Upper and lower bounds for quadratic functionals. *Proceedings of the Symposium on Spectral Theory and Differential Problems.* Oklahoma A. & M. College (1951) p. 279.
3. Diaz, J. B. Estimation of torsional rigidity and other physical quantities. *Proceedings of the First U. S. National Congress of Applied Mechanics.* Chicago, (1951).
4. Diaz, J. B. and A. Weinstein. Schwarz's inequality and the methods of Rayleigh-Ritz and Trefftz. *Journal of Mathematics and Physics* 3 (1947) p. 113.
5. Diaz, J. B. and A. Weinstein. The torsional rigidity and variational methods. *American Journal of Mathematics* 70 (1948) p. 107.

6. Polya, G. and G. Szegö. *Isoperimetric inequalities in mathematical physics*. Princeton, 1951.
7. Saint-Venant, B. Mémoire sur la torsion des prismes. *Mémoires présentés par divers savants à l'Académie des Sciences de l'Institut Impérial de France* 14 (1856) p. 233.
8. Weber, C. Bestimmung des Steifigkeitswertes von Körpern durch zwei Näherungsverfahren. *Zeitschrift fur Angewandte Mathematik*. 11 (1931) p. 244.
9. Weinel, E. Das Torsionsproblem fur den exzentrischen Kreisrung. *Ingenieur Archive* 3 (1932) p. 67.

Institute for Fluid Dynamics and Applied Mathematics
University of Maryland

(Received April 23, 1952)

AN ABRIDGED BLOCK METHOD FOR THE SOLUTION OF THE DIRICHLET PROBLEM FOR THE LAPLACE DIFFERENCE EQUATION

By Charles Saltzer*

1. Introduction. For simplicity the method given in this note will be presented for the case of two dimensional networks with square meshes. However, the method is general and the extension to more general meshes and to spaces of higher dimensions will be indicated. This method is based on the use of the formula for the solution of the Dirichlet net problem for a rectangular region. (The corresponding formula for the three dimensional case was given by N. Wiener and H. B. Phillips [1]. A special case of the formula for two dimensions was developed by W. H. McCrea and F. J. W. Whipple [2] and the general formula for the two dimensional case was given by M. A. Hyman [3].) (See also [8] and [9])

The advantage of this method is that the number of equations which arise in the usual procedure is, in general, considerably reduced, and in some cases the number becomes so small that the system of equations can be directly inverted.

If the equations which occur in this method are solved by relaxation or iteration, the convergence of these procedures will be seen to be a direct consequence of the convergence proofs given by J. B. Diaz and R. C. Roberts [4].

In addition, the formula for the solution for a rectangular region, equation (6), is given in a slightly modified form which simplifies the application of the method.†

2. Formulation of the problem. Consider the set of points in the xy-plane whose coordinates are (ph, qh) $(p, q = 0, \pm 1, \pm 2, \cdots)$. We say that two points are adjacent if the distance between them is h. A connected net, N, is a finite subset of the above set which has the property that for any two points of the net there is at least one sequence of points of the net whose initial and terminal points are the given points and the distance between consecutive points of the sequence is h. We shall consider only connected nets. A point of the net is an interior point if it is adjacent to four points of the net, and, it is a boundary point of it is adjacent to at least one inner point, but is not an inner point. A point of a connected net which is neither a boundary point or an inner point will be called an irregular point of the net. A rectangular net, R, is the set of points (ph, qh), $a \leqq p \leqq b$, $c \leqq q \leqq d$ where a, b, c, d, p and q are integers. A rectangular net whose four irregular points have been deleted will also be regarded as a

* Presented to the Am. Math. Soc. Dec. 28, 1952.

† When this note was presented, Dr. P. C. Rosenbloom informed me that Mr. Leon Brown had obtained results similar to those of this note and is engaged in the calculation of tables to facilitate the application of this method. See also [7].

rectangular net. A function defined on a net is lattice-harmonic at a point if the value of the function at this point is the mean of its values on the four adjacent points.

For a given net the Dirichlet problem for the Laplace difference equation consists of finding a function which has prescribed values on the boundary points of the lattice and is lattice-harmonic at all the interior points.

3. The block method. We consider first a special case, as in figure 1, where the net, N, consists of the sum of two rectangular nets situated so that the right boundary of one rectangle L, and the left boundary of the other rectangle R have three or more boundary points in common. We also assume that the lower sides of the rectangles are not collinear, and we choose our coordinate axes so that the y-axis passes through the common points, and the x-axis through the upper of the lower sides of the rectangles. If there are n common points $a_1, a_2, \cdots a_n$, which are also interior points of N, their coordinates are $(0, h)$, $(0, 2h), \cdots, (0, nh)$. (The point $(0, 0)$ is an irregular point of R). Let the point

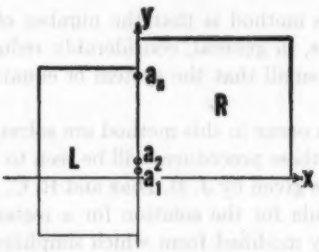


FIGURE 1

set B: $b_1, b_2, \cdots b_r$ be those boundary points of L which are not boundary points of R; let the point set C: $c_1, c_2, \cdots c_s$ be those boundary points of R which are not boundary points of L. We also assume that the upper sides are not collinear. In this case the boundary points of N are $B + C$. In addition let $U(ph, qh)$ be a function defined on N which is harmonic in N and whose boundary values on B are $U = V_k$ at $b_k (k = 1, 2, \cdots r)$, $U = W_k$ at $c_k (k = 1, 2, \cdots s)$; and finally, let $U = U_k$ at $a_k (k = 1, 2, \cdots n)$. By section (5), U can be represented in L by

$$U(-ph, qh) = \sum_{k=1}^{n} \sigma_{pqk} U_k + \sum_{k=1}^{r} \sigma_{pqk+n} V_k \qquad (1)$$

where σ_{pqk} $(k = 1, 2, \cdots, n + r)$ are constants defined in section (5), and in R by

$$U(ph, qh) = \sum_{k=1}^{n} \mu_{pqk} U_k + \sum_{k=1}^{s} \mu_{pqk+n} W_k \qquad (2)$$

where $\mu_{pqk} (k = 1, 2, \cdots n + s)$ are constants defined in section (5). If we write these equations for the points (h, qh) and $(-h, qh)$ $(q = 1, 2, \cdots n)$ we get

$$U(h, qh) = \sum_{k=1}^{n} \alpha_{qk} U_k + \xi_q, \qquad U(-h, qh) = \sum_{k=1}^{n} \beta_{qk} U_k + \eta_q$$

where

$$\alpha_{qk} = \mu_{1qk}, \qquad \beta_{qk} = \sigma_{1qk},$$
$$\xi_q = \sum_{k=1}^{s} \mu_{1qk+n} W_k, \qquad \eta_q = \sum_{k=1}^{r} \sigma_{1qk+n} V_k.$$

Applying the condition that U be harmonic at $a_1, a_2, \cdots a_n$, using the above representations for the values of U at the adjacent inner points, and noting that $U(0, qh) = U_q$, yields

$$4U_q = U_{q+1} + U_{q-1} + \sum_{k=1}^{n} (\alpha_{qk} + \beta_{qk}) U_k + (\xi_q + \eta_q)$$
$$(q = 2, 3, \cdots n - 1), \qquad (3)$$

with two similar equations for $q = 1$ and $q = n$. If the values of U are known only at the boundary points of N then this system of n equations is a system of n linear simultaneous equations for the n unknowns U_k, since α_{qk}, β_{qk}, ξ_q and η_q can be calculated from V_k and W_k. Further, if U_k $(k = 1, 2, \cdots n)$ are known then (1) and (2) permit the calculation of U at any interior point.*

4. General two dimensional region. The above procedure can be applied to a net decomposed into a finite number of abutting or overlapping rectangles and any finite net can be decomposed in this way. In any event, by using (1) for each of the rectangles and then applying the condition that the desired function be harmonic at each point which is both a boundary point of a rectangle and an inner point of the net,† we get a system of as many equations as unknowns. Once these are calculated, equation (1) permits the calculation of U as before. In the case where a rectangle has only one inner point, equation (3) reduces to the condition that the function be the mean of its values at the adjacent points. The proof of the existence and uniqueness of the solution is an immediate consequence of the classical theorems for this problem [5] since $V_k = W_k = 0$ for all k implies that $U_k = 0$ for $k = 1, 2, \cdots n$.

If the number of equations in the above system is not small enough to be solved directly it may be noted that these equations are in a form suitable for iteration or relaxation. The convergence of either of these processes is assured by the theorems of J. B. Diaz and R. C. Roberts [4].

5. The solution for a rectangle and the derivation of equations (1) and (2). The usual procedure on calculating the solution for a rectangle is to note that $U = e^{i\alpha ph + \beta qh}$ is lattice-harmonic at any point if $\cos \alpha h + \cosh \beta h = 2$. This condition is obtained by the substitution of U in the Laplace difference equation. If the rectangle is the net for which $0 \leqq p \leqq m$, $0 \leqq q \leqq n$, $mh = a$, $nh = b$, then the solutions which are zero on both sides and the top are

$$U(ph, qh) = \sum_{k=1}^{m-1} B_k \frac{\sinh \beta_k (b - qh)}{\sinh \beta_k b} \sin \frac{\pi kph}{a} \qquad (4)$$

where $\cosh \beta_k h = 2 - \cos(\pi kh/a)$ and $B_1, B_2 \cdots B_{m-1}$ are arbitrary constants.

* This method is a finite difference analogue of the method given in [10].

† For overlapping rectangles a slight modification involving irregular points can readily be made.

If, in addition, we require that $U'(ph, 0) = f_p$ $(p=1, 2, \cdots m-1)$, U will satisfy this condition [6] provided we set

$$B_k = \sum_{r=1}^{m-1} f_r S_{kr} \quad \text{where} \quad S_{kr} = (2/m) \sin (kr\pi h/a)$$

By substitution in equation (4) we get

$$U'(ph, qh) = \sum_{r=1}^{m-1} C'_{pqr} f_r \tag{5}$$

where

$$C'_{pqr} = \sum_{k=1}^{m-1} S_{kr} \sin (\pi kph/a) \sinh \beta_k (b - qh)/\sinh \beta_k b.$$

If we use $\sinh \beta_k qh$ instead of $\sinh \beta_k(b-qh)$ and if $U''(ph, nh) = g_p$ $(p = 1, 2, \cdots m-1)$ and $U'' = 0$ at all other boundary points, then

$$U''(ph, nh) = \sum_{r=1}^{m-1} C''_{pqr} g_r$$

where

$$C''_{pqr} = \sum_{k=1}^{m-1} S_{kr} \sin(\pi kph/a) \sinh \beta_k qh/\sinh \beta_k b.$$

Interchanging the roles of p and q in (5) gives

$$U'''(ph, qh) = \sum_{r=1}^{n-1} C'''_{pqr} \phi_r$$

where

$$C'''_{pqr} = \sum_{k=1}^{n-1} (T_{kr} \sin (\pi kqh/b) \sinh \beta'_k (a-ph)/\sinh \beta'_k a),$$

$$T_{kr} = (2/n) \sin (kr\pi h/b), \qquad \cosh \beta'_k h = 2 - \cos (\pi kh/b),$$

and $U'''(0, qh) = \phi_q$ $(q = 1, 2, \cdots n-1)$, $U''' = 0$ at all other boundary points. In the same way, if $U^{\text{iv}}(mh, qh) = \psi_q$ $(q = 1, 2, \cdots n-1)$ and $U^{\text{iv}} = 0$ at all other boundary points then

$$U^{\text{iv}}(ph, qh) = \sum_{r=1}^{n-1} C^{\text{iv}}_{pqr} \psi_r$$

where

$$C^{\text{iv}}_{pqr} = \sum_{k=1}^{n-1} (T_{kr} \sin (\pi kqh/b) \sinh \beta'_k ph/\sinh \beta'_k a).$$

Hence if $U = U' + U'' + U''' + U^{\text{iv}}$ then

$$U(ph, qh) = \sum_{r=1}^{m-1} (C'_{pqr} f_r + C''_{pqr} g_r) + \sum_{r=1}^{n-1} (C'''_{pqr} \phi_r + C^{\text{iv}}_{pqr} \psi_r) \tag{6}$$

is that lattice-harmonic function which assumes the values f_r on the lower boundary points of the rectangle, g_r on the upper boundary points, ϕ_r on the left boundary points, and ψ_r on the right boundary points. It may be noted that $U = 0$ at the irregular points of the net.

To get eq. (1) we set $\phi_k = U_k$ and $C'_{pqk} = \sigma_{pqk}$ $(k = 1, 2, \cdots n-1)$ and proceed around the boundary identifying the values of σ and V with appropriate values of g, ϕ and ψ. Equation (2) is obtained in a similar way.

For the case of a rectangular mesh of width h and height k, the formula corresponding to equation (6) can be obtained by substituting $e^{i\alpha ph+\beta qk}$ in the corresponding Laplace difference equation and proceeding as above.

6. Higher dimensional spaces. The procedure described in section (4) can be applied to nets in spaces of any number of dimensions. The formula for the three dimensional case corresponding to eq. (6) is given in [1]. The method used to obtain this formula is readily extended to any number of dimensions.

BIBLIOGRAPHY

1. Wiener, N. and Phillips, H. B.: Nets and the Dirichlet problem, J. Math. and Phys. **1**, 105–124 (1922).
2. McCrea, W. H. and Whipple, F. J. W.: Random paths in two and three dimensions, Proc. Roy Soc. of Edinburgh, **60**, 281–298 (1939–1940).
3. Hyman, M. A.: Non-iterative numerical solutions of boundary value problems, U. S. Naval Ordnance Laboratory, White Oak Maryland, Report 1813.
4. Diaz, J. B., and Roberts, R. C.: On the numerical solution of the Dirichlet problem for Laplace's difference equation. Quart. of Appl. Math. **9**, 4 (1952). (See method II p. 359).
5. Feller, W., and Tamarkin, J. D.: Partial differential equations, Brown University, 1941 (chapter V).
6. Jackson, D.: The theory of approximation, New York (1930) (chapter IV).
7. Moskovitz, D.: The numerical solution of Laplace's and Poisson's equations, Quart. of Appl. Math. **2** (1944), pp. 148–163.
8. Le Roux, J.: Sur le problème de Dirichlet, Jnl. de Math. Pures et Appliquees, **10** (1914), pp. 189–230.
9. Perron, O.: Explizite Lösung einer gewissen partiellen Differenzengleichung bei gegebenen Randwerten auf einem Rechteck, S.-B. Heidelberg Akad. Wiss. (1934), pp. 40–48.
10. Epstein, B.: A method for the solution of the Dirichlet problem for certain types of domains. Quart. of Appl. Math. **6** (1948), pp. 301–317.

Case Institute of Technology

(Received April 3, 1952)

NOTE ON THE FITTING OF POLYNOMIALS TO EQUALLY-SPACED OBSERVATIONS

By P. G. Guest

In the fitting of polynomials to equally-spaced observations it is customary to employ orthogonal moments calculated from tables of orthogonal polynomials [1]. However, when the observations are not equally-spaced power moments must be used, and two papers [2, 3] have recently appeared in which the connection between the methods of fitting by power moments and by orthogonal moments is discussed. In the present note it will be shown that for the equally-spaced case the method using power moments is at least as convenient as the method using tables of orthogonal polynomials.

TABLE 1

$S_{00} = n$	$\beta_{kk} = 1$
$S_{11} = n(n^2 - 1)/12$	$\beta_{kj} = 0$ if $j + k$ is odd
$S_{22} = n(n^2 - 1)(n^2 - 4)/180$	$\beta_{20} = -(n^2 - 1)/12$
$S_{33} = n(n^2 - 1)(n^2 - 4)(n^2 - 9)/2800$	$\beta_{31} = -(3n^2 - 7)/20$
$S_{44} = n(n^2 - 1)(n^2 - 4)(n^2 - 9)(n^2 - 16)/44100$	$\beta_{42} = -(3n^2 - 13)/14$
	$\beta_{40} = +3(n^2 - 1)(n^2 - 9)/560$

TABLE 2

$S_2 = \frac{1}{2}S_{11}$

$A_{13} = -\alpha_{20}$, $A_{33} = S_{22}/S_{11}$

$A_{24} = -\alpha_{31}$, $A_{44} = S_{33}/S_{11}$

$A_{15} = -\alpha_{40}$, $A_{35} = -\alpha_{42}$, $A_{55} = S_{44}/S_{11}$, $A_{53} = -\alpha_{42}S_{22}/S_{11}$

Let M_r denote the power moment $\sum_\epsilon y(\epsilon)\epsilon^r$ and b_{pj} the estimate of the coefficient of ϵ^j in the fitted polynomial of degree p. The independent variable ϵ is chosen so that its origin lies at the centre of the range of observations and the interval between successive observations is unity. The relations required for fitting by power moments are [2]

$$M'_k = \sum_{r=0}^{k} \beta_{kr} M_r \tag{1}$$

$$a_k = M'_k / S_{kk} \tag{2}$$

$$b_{pj} = \sum_{k=j}^{p} \beta_{kj} a_k . \tag{3}$$

β_{kj} and S_{kk} are functions of n, the number of observations. The expressions for these quantities in terms of n are given in Table 1 for values of k up to 4. A table giving numerical values of β_{kj} and S_{kk} for values of n from 6 to 104 has been deposited in the Unpublished Mathematical Tables file of the periodical Mathematical Tables and other Aids to Computation [4].

Model form 1 — n even

$M_j = \Sigma c^j(y_+ + (-)^j y_-)$				
$3c^3$	$2c$	y_+	y_-	$4c^4$
1	1	5	13	1
27	3	6	8	9
125	5	10	6	25
343	7	8	5	49
729	9	10	10	81
1331	11	13	13	121
2197	13	10	9	169
3375	15	3	10	225
4913	17	7	5	289
6859	19	16	5	361
9261	21	29	2	441
12167	23	37	1	529
15625	25	38	8	625
19683	27	50	3	729
24389	29	74	6	841
29791	31	22	22	961
35937	33	19	36	1089
42875	35	44	30	1225
50653	37	35	20	1369
59319	39	15	21	1521
68921	41	15	24	1681
79507	43	18	28	1849
91125	45	15	45	2025
103823	47	10	52	2209
117649	49	6	57	2401
132651	51	4	56	2601
148877	53	0	48	2809
166375	55	3	55	3025
185193	57	1	73	3249
205379	59	3	65	3481
226981	61	7	67	3721
250047	63			3969
274625	65			4225
300763	67			4489
328509	69			4761
357911	71			5041
389017	73			5329
421875	75			5625
456533	77			5929
493039	79			6241
531441	81			6561

S_{00} 62

S_{11} 19855.5

S_{22} 5,083008

S_{33} 1353,143008

β_{20} −320.25

β_{31} −576.25

$a_j = M'_j/S_{jj}$	Σy^2 54038
$M_0 = M'_0 + 1336$	$a_0M'_0$ 28789
$a_0 + 21.54839$	Σv_0^2 25249
$M_1 = M'_1 - 10143$	$a_1M'_1$ 5181
$a_1 - 0.5108408$	Σv_1^2 20068
$M_2 + 573404$	$a_2M'_2$ 4168
$+\beta_{20}M_0 - 427854$	Σv_2^2 15900
$= M'_2 + 145550$	(s_2)
$a_2 + 0.02863462$	
$M_3 - 9086574.75$	$a_3M'_3$ 8386
$+\beta_{31}M_1 + 5844903.75$	Σv_3^2 7514
$= M'_3 - 3241671$	(s_3) 11.38
$a_3 - 0.002586832$	

$a_3 = b_{33} - 0.002586832$

$\beta_{31}a_3 + 1.490662$

$+a_1 - 0.5108408$

$= b_{p1} + 0.979821$

$a_2 = b_{p2} + 0.02863462$

$\beta_{20}a_2 - 9.17024$

$+a_0 + 21.54839$

$= b_{p0} + 12.37815$

$M_j = \Sigma \epsilon^j (y_+ + (-)^j y_-)$							
$8\epsilon^3$	2ϵ	y_+ y_-	$4\epsilon^2$				
571787	83		6889				
614125	85		7225	$s[b_{pj}] = \rho_{pj}(O, n) s_p / n^{j+\frac{1}{2}}$			
658503	87		7569				
704969	89		7921	j	$\rho_{pj}(0)$	$n^{j+\frac{1}{2}}$	$s[b_{pj}]$
753571	91		8281	0	*1.50*	*7.874*	*2.17*
804357	93		8649				
857375	95		9025	1	*8.66*	*488.2*	*0.202*
912673	97		9409				
970299	99		9801	2	*13.4*	*5027 × 10*	*0.00504*
1030301	101		10201				
1092727	103		10609	3	*53.0*	*1877 × 10²*	*0.000321*

Note that the values found by multiplying the columns must be divided by 2^j to give M_j.

The use of power moments is advocated because it enables a very simple Model Form to be devised by means of which the calculations can be rapidly carried out. In the method using tables of orthogonal polynomials the orthogonal moments M'_k are obtained directly. Thus the only extra calculations to be performed in the method using power moments are those given by equation (1), and these are very rapidly done, as will be seen in the example worked out below. But the orthogonal polynomials $\xi'_j(\epsilon)$ are functions of n as well as of ϵ, while the powers ϵ^j are of course independent of n. Thus the powers may be printed in a permanent Model Form which can be used for all values of n, and this greatly simplifies the calculation of the moments.

Model Form 1 shows the proposed scheme applied to the example of Anderson and Houseman [1]. The 62 observations are entered in order in the y columns, starting from the lower portion of the left-hand column and working up this column and down the right-hand column. The moments are then evaluated on a calculating machine, and $\sum y^2$ is also determined. The coefficients are then calculated on the right-hand side of the Model Form.

The estimated error s_p of a single observation may be obtained in the usual way from the formula

$$s_p = [\sum v_p^2/(n - p - 1)]^{\frac{1}{2}}, \tag{4}$$

and the errors of the coefficients from the tabulated weight-functions [5] ρ_{pj} (or directly from the values β_{kj}, S_{kk} [2]).

The Model Form has been designed to fit on to a foolscap page so that mimeographed copies can be made. Separate Model Forms are required for the cases in which n is odd and the cases in which n is even, since ϵ takes integral values when n is odd and half-integral values when n is even. If the polynomial of the fourth degree is to be calculated (and only very rarely is this required), a separate sheet containing values of ϵ^4 is placed alongside the column of values y to obtain the moment M_4.

The present method is closely related to the method using power moments described by Nielsen and Goldstein [6]. Their method is based essentially on coefficients α_{kj} (c.f. reference [2]) such that

$$M'_k = M_k + \sum_{r=0}^{k-1} \alpha_{kr} M'_r \qquad (1')$$

$$a_k = M'_k / S_{kk} \qquad (2')$$

$$b_{pj} = a_j + \sum_{k=j+1}^{p} \alpha_{kj} b_{pk}. \qquad (3')$$

However, their notation is complicated by a division by S_{11}. Table 2 shows the relationships between the quantities they tabulate and the quantities in the above equations. An extra quantity A_{kj} is required in their method because equations (1′) and (2′) are combined into a single equation.

The advantage of the present method lies principally in the simplicity of the notation and the convenience of the Model Form. The standard errors can also be obtained very quickly. The quantities β_{kj}, S_{kk} by means of which the fitting is accomplished are of interest in themselves, β_{kj} being the coefficient of ϵ^j in the orthogonal polynomial $\xi_k(\epsilon)$ and S_{kk} being the sum $\sum_\epsilon \xi_k^2(\epsilon)$. The use of the quantities β_{kj} rather than the quantities α_{kj} or A_{kj} enables the coefficients b_{pj} to be expressed in terms of the coefficients a_j. Since the latter are independent of p, this is a slight advantage if the least-squares polynomials of different degrees are required.

REFERENCES

[1] R. L. Anderson and E. E. Houseman, "Tables of orthogonal polynomial values extended to $N = 104$," Iowa State College Agric. Exp. Station, Research Bulletin 292, 1942.

[2] P. G. Guest, "Orthogonal polynomials in the least squares fitting of observations," Phil. Mag. [7], **41**, 1950, pp. 124–137.

[3] J. G. Hayes and T. Vickers, "The fitting of polynomials to unequally-spaced data," Phil. Mag. [7], **42**, 1951, pp. 1387–1400.

[4] P. G. Guest, "Tables of certain functions occurring in the fitting of polynomials to equally-spaced observations," deposited in UMT file, Math. Tables Aids Comp., **6**, no. 37, 1952, p. 40.

[5] P. G. Guest, "Estimation of the error at a point on a least-squares curve," Australian J. Sci. Res., Ser. A, **3**, 1950, pp. 173–182. "Estimation of the errors of the least-squares polynomial coefficients," ibid., pp. 364–375.

[6] K. L. Nielsen and L. Goldstein, "An algorithm for least squares," J. Math. and Phys., **26**, 1947, pp. 120–132.

University of Sydney, Australia

(Received June 4, 1952)

ON THE TRANSONIC FLOW PAST A FINITE WEDGE*

By L. Trilling and K. Walker, Jr.

In a recent paper (1), J. D. Cole investigated the steady plane flow of a perfect ideal gas past a thin finite wedge parallel to the free stream direction at a high subsonic velocity. The transonic equations were assumed to be valid, and the problem was formulated and solved approximately in the hodograph plane where the coordinate normal to the free stream direction (or the stream function) satisfies the Tricomi partial differential equation (2) with the velocity perturbations from sonic velocity parallel to the free stream direction as independent variables.

While that investigation gave one of the earliest solutions to a specific boundary value problem in transonic flow, and the drag it predicted agreed quite well with the measurements of Liepmann and Bryson (3, 4) and Griffith (5, 6), Cole's solution did not satisfy all the boundary conditions of the problem. In particular, the condition that the flow turns the shoulder corner of the wedge (point B on Fig. 1) by a Prandtl Meyer expansion was replaced by the much simpler approximate condition that the sonic line is vertical. In this way, Cole avoided discussing the structure and flow pattern of the local supersonic region, while satisfying all the conditions in the subsonic field, and constructing the correct hodograph singularity to represent the free stream. He felt that since the flow outside a stream tube situated some distance away from the wedge is subsonic, and since the drag of the wedge is defined by the force on such a stream tube, a good approximation of the drag could be obtained if the flow far from the wedge were well represented, even though the approximation was poor over part of the surface of the wedge.

It is the purpose of the present paper to construct a solution of the Tricomi equation which satisfies all the boundary conditions of the half-wedge problem, in order both to evaluate the accuracy of Cole's simpler solution, and to gain more detailed information on the shape and structure of the local supersonic region. Consider the flow past a wedge symmetric about the x axis; with a semi wedge angle δ and unit length, and its leading edge at the origin (Fig. 1). The free stream Mach number is M.

As Cole and others have shown, the equations of motion in the hodograph plane are:

$$uy_v - x_u = 0 \tag{1a}$$

$$y_v - x_u = 0 \tag{1b}$$

where x, y are the position coordinates, made dimensionless by division by the wedge length, and the velocity perturbations u, v are defined as follows:

$$u = [(\gamma + 1)/a^*]u'; \qquad v = [(\gamma + 1)/a^*]v' \tag{2}$$

* This paper is based on work done at the Massachusetts Institute of Technology for the Wright Air Development Center, United States Air Force, under contract AF33(038)-22184 and on parts of a thesis submitted in partial fulfillment of the requirements for the degree of Master of Science by the junior author.

Here a^* is the velocity of sound at $M = 1$, γ the ratio of specific heats and u', v' are the parallel and normal velocity perturbations from a^* directed along the x axis.

The boundary conditions require that the flow far from the wedge have free stream velocity, that the velocity component normal to the wedge surface vanish

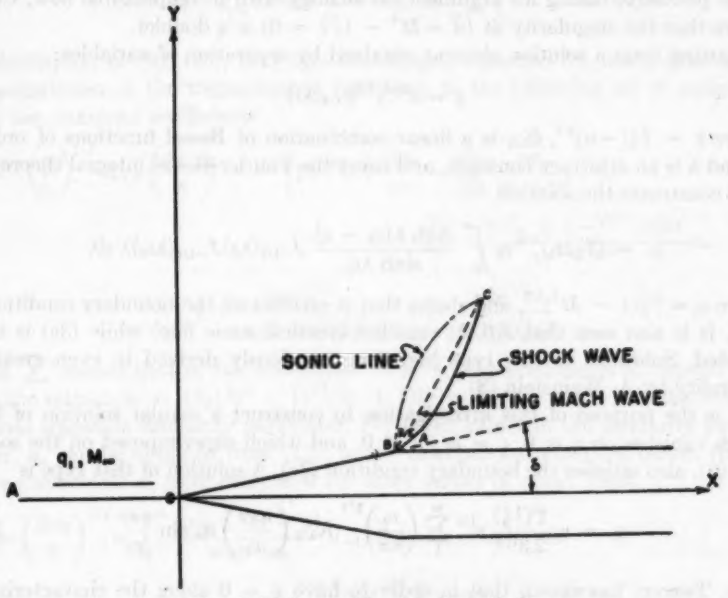


FIG. 1. Wedge problem in physical plane

and that the flow turn the corner at $x = 1$ by a Prandtl Meyer expansion. In the physical plane, this means

$$\lim_{\substack{x\to\infty \\ y\to\infty}} u(x, y) = M^2 - 1; \quad \lim_{\substack{x\to\infty \\ y\to\infty}} v(x, y) = 0 \tag{3a}$$

$$v(x, 0) = \begin{cases} 0 & (-\infty < x < 0) \\ (\gamma + 1)\delta = \delta_0 & (0 < x < 1) \\ 0 & (1 < x < \infty) \end{cases} \tag{3b}$$

$$v(1, 0) = v_0 - \tfrac{2}{3}u^{3/2} \tag{3c}$$

with the further condition that the origin is a stagnation point so that

$$\lim_{\substack{x\to 0 \\ y\to 0}} u = -\infty \tag{3d}$$

In the hodograph, boundary conditions (3b, 3d) give

$$y(u, v_0) = 0, \qquad y(u, 0) = 0 (u \neq M^2 - 1), \qquad (u < 0) \quad (4a)$$

$$\lim_{\substack{v \to 0 \\ u \to M^2 - 1}} y(u, v) = \infty$$

More precisely, basing his argument on analogy with incompressible flow, Cole shows that the singularity at ($u = M^2 - 1$, $v = 0$) is a doublet.

Starting from a solution element obtained by separation of variables:

$$y = e^{\pm\lambda v} z^{1/3} \mathfrak{C}_{1/3}(\lambda z) \tag{5}$$

where $z = \frac{2}{3}(-u)^{3/2}$, $\mathfrak{C}_{1/3}$ is a linear combination of Bessel functions of order $\frac{1}{3}$ and λ is an arbitrary constant, and using the Fourier Bessel integral theorem, Cole constructs the solution

$$y_s = (\tfrac{2}{3} z z_1)^{1/3} v_0 \int_0^\infty \frac{\sinh \lambda(v_0 - v)}{\sinh \lambda v_0} J_{-1/3}(\lambda z) J_{-1/3}(\lambda z_1) \lambda \, d\lambda$$

where $z_1 = \frac{2}{3}(1 - M^2)^{3/2}$, and shows that it satisfies all the boundary conditions (4a). It is also seen that $x(0, v)$ vanishes (vertical sonic line) while (3c) is not satisfied. Solutions of this type had been previously derived in even greater generality by A. Weinstein (8).

It is the purpose of this investigation to construct a regular solution of (1) which vanishes on $v = 0$, $v = v_0$, $u < 0$, and which superimposed on the solution (6), also satisfies the boundary condition (3c). A solution of that type is

$$y_c = v_0 \frac{\Gamma(\frac{1}{3})}{2.3^{1/6}} z^{1/3} \sum_1^\infty \left(\frac{v_0}{n\pi}\right)^{1/3} K_{1/3}\left(\frac{n\pi z}{v_0}\right) a_n \sin \frac{n\pi v}{v_0} \tag{7}$$

Now, Tricomi has shown that in order to have $y = 0$ along the characteristic through $(0, v_0)$, it is necessary and sufficient to satisfy the following integral equation along the sonic line:

$$y(0, v) = -\frac{3^{2/3}[\Gamma(\frac{1}{3})]^3}{4\pi^2} \int_{v_0}^v \frac{\partial y}{\partial u}(0, \alpha)(v - \alpha)^{-1/3} \, d\alpha \tag{8}$$

The problem is therefore to find the coefficients a_n which will satisfy this relation.

When the left hand side of (8) is expanded into a Fourier sine series in the half interval $0 < v < v_0$, it becomes:

$$y(0, v) = \left(\frac{4z_1}{3}\right)^{1/3} \frac{v_0}{\Gamma(\frac{2}{3})} \int_0^\infty \frac{\sinh \lambda(v_0 - v)}{\sinh \lambda v_0} J_{-1/3}(\lambda z_1) \lambda^{2/3} \, d\lambda$$

$$+ 3^{-1/3} \frac{\Gamma(\frac{1}{3})}{\Gamma(\frac{2}{3})} \sum_1^\infty \left(\frac{v_0}{n\pi}\right)^{2/3} a_n \sin \frac{n\pi v}{v_0} \tag{9}$$

$$= \frac{3^{-1/3}}{\Gamma(\frac{2}{3})} \sum_1^\infty \left[\left(\frac{32 n^2 \pi^2 z_1}{v_0^2}\right)^{1/3} K_{1/3}\left(\frac{n\pi z_1}{v_0}\right) + \frac{\Gamma(\frac{1}{3})}{v_0}\left(\frac{v_1}{n\pi}\right)^{2/3} a_n\right] \sin \frac{n\pi v}{v_0}$$

The evaluation of the Fourier coefficient of y_s above was carried out with the aid of Ref. 7, p. 434. Similarly

$$\frac{\partial y}{\partial u} = \frac{1}{v_0} \sum_1^\infty a_n \sin \frac{n\pi v}{v_0} \tag{10}$$

$$v^{-1/3} = v_0^{-1/3} \sum_1^\infty c_n \sin \frac{n\pi v}{v_0} \qquad c_n = \frac{2}{\pi^{2/3}} \int_0^\pi \theta^{-1/3} \sin n\theta \, d\theta \tag{11}$$

Substitution of (9), (10), (11) into the integral equation (8) leads, after some manipulation of the trigonometric functions, to the following set of equations for the unknown coefficients

$$\zeta n^{2/3} \left(\frac{z_1}{v_0}\right)^{1/3} K_{1/3}\left(\frac{n\pi z_1}{v_0}\right) = -\left\{\xi n^{-2/3} + \frac{3c_n}{4n} - \sum_{p=1}^{\infty}{}' \frac{pc_p}{n^2 - p^2}\right\} a_n + \sum_1^\infty{}' \frac{-mC_n + (-)^{n+m} nCm}{m^2 - n^2} a_m \tag{12}$$

$$\zeta = \frac{(4\pi^2)^{4/3}}{\sqrt{3}\,[\Gamma(\tfrac{1}{3})]^2} \qquad \xi = \frac{2\pi^{4/3}}{\sqrt{3}\,\Gamma(\tfrac{1}{3})}$$

and $\sum'$ means that in the summation $m \neq n$. The coefficients a_n depend only on the ratio $z_1/v_0 = (\tfrac{2}{3}(M^2 - 1)^{3/2}/(\gamma + 1)\delta) = \tfrac{2}{3}\xi_0^{3/2}$, where ξ_0 is the conventional transonic similarity parameter. For given values of the similarity parameter, a_n can be obtained from (12); in the calculations presented below, thirty-five coefficients a_n were computed for three values of z_1/v_0. Then,

$$y = \left(\frac{2zz_1}{3}\right)^{1/3} v_0 \int_0^\infty \frac{\sinh \lambda\,(v_0 - v)}{\sinh \lambda v_0} J_{-1/3}(\lambda z)\, J_{-1/3}(\lambda z_1) \lambda^2 \, d\lambda + \frac{\Gamma(\tfrac{1}{3}) z^{1/3}}{3^{1/6} \pi^{4/3} v_0^{2/3}} \sum_1^\infty n^{-1/3} a_n K_{1/3}\left(\frac{n\pi z}{v_0}\right) \sin \frac{n\pi v}{v_0} \tag{13}$$

Substituting this into the irrotationality equation (1b), one obtains, making use of the fact that $x(0, 0) = 1$

$$x = 1 + \frac{v_0}{\pi} \sum_1^\infty a_n \left[\frac{1}{n} - \frac{\Gamma(\tfrac{1}{3}) n^{-1/3}}{3^{1/6} \pi^{1/3} v_0^{2/3}} z^{2/3} K_{2/3}\left(\frac{n\pi z}{v_0}\right) \cos \frac{n\pi v}{v_0}\right] - z^{2/3} z_1^{1/3} v_0 \int_0^\infty \frac{\cosh \lambda (v_0 - v)}{\sinh \lambda v_0} J_{2/3}(\lambda z) J_{-1/3}(\lambda z_1) \lambda \, d\lambda \tag{14}$$

Equations (13), (14) represent the required solution, from which the structure of the flow field must be obtained. The integrals in (13), (14) converge when $u \leqq 0$ and $v - \tfrac{2}{3}u^{3/2} \geqq 0$ so that this representation is valid in the entire region for which the specified boundary conditions define a unique solution (flow up to the limiting Mach line). Another representation of the integrals in (13), (14) as a series of Legendre functions is needed beyond the limiting Mach line $v = \tfrac{2}{3}u^{3/2}$. It may be obtained as follows from Ref. 7, p. 389.

$$y_s = \left(\frac{2zz_1}{3}\right)^{1/3} v_0 \int_0^\infty \frac{\sinh \lambda(v_0 - v)}{\sinh \lambda v_0} J_{-1/3}(\lambda z) J_{-1/3}(\lambda z_1) \lambda \, d\lambda$$

$$= \left(\frac{2zz_1}{3}\right)^{1/3} v_0 \int_0^\infty \left\{ \sum_{n=0}^{\infty} e^{-\lambda(v+2nv_0)} - \sum_{n=0}^{\infty} e^{-\lambda[2(n+1)v_0 - v]} \right\} J_{-1/3}(\lambda z) J_{-1/3}(\lambda z_1) \lambda \, d\lambda \quad (13a)$$

$$= -\left(\frac{2zz_1}{3}\right)^{1/3} \frac{v_0}{\pi} \frac{1}{\sqrt{zz_1}} \frac{\partial}{\partial v} \sum_{n=0}^{\infty} Q_{-5/6}\left[\frac{(v + 2nv_0)^2 + z^2 + z_1^2}{2zz_1}\right] - Q_{-5/6}\left[\frac{[2(n+1)v_0 - v]^2 + z^2 + z_1^2}{2zz_1}\right]$$

and a similar formula for $x_s(u, v)$.

The shape of the sonic line is obtained by setting $z = 0$ in the formulas (13), (14) and carrying out the necessary limiting processes: thus

$$x(0, v; z_1) = 1 + \frac{1}{\pi} \sum_1^\infty \frac{a_n}{n} \left(1 - \cos \frac{n\pi v}{v_0}\right)$$

$$y(0, v; z_1) = \frac{3^{-1/3}}{\Gamma(\frac{2}{3})} \sum_1^\infty \left[\frac{2}{z_1^{1/3}} \left(\frac{2n\pi z_1}{v_0}\right)^{2/3} K_{1/3}\left(\frac{n\pi z_1}{v_0}\right) + \frac{\Gamma(\frac{1}{3})}{v_0} \left(\frac{v_0}{n\pi}\right)^{2/3} a_n\right] \sin \frac{n\pi v}{v_0} \quad (15)$$

From these equations the sonic line was computed for three values of the transonic similarity parameter, and the results are shown on Figure 3.

It follows from the general transonic similarity theory, and from the form of equations (15), (16) that x, $yv_0^{1/3}$ are functions of z_1/v_0 only. Comparisons of x, $yv_0^{1/3}$ with Bryson's measurements (4) on a 10° wedge were made at several values of z_1/v_0. It was found that when z_1/v_0 was larger than 0.5, no supersonic region could be detected on the interferograms, either because of lack of resolution or because of the smoothing out of the corner flow by viscous effects. Agreement between experiment and theory was found to be fair in the proper transonic region (see Fig. 3). It is the same order of agreement as between experiment and Drebinger's theory for slightly supersonic flow (Ref. 4).

The drag coefficient is found by integrating the pressure coefficient along the wedge sides, which gives:

$$C_D = 2\delta \int_0^1 c_p \, dx = \frac{4\sigma}{a^*} \int_0^1 (u_1 - u) \, dx = \frac{4\delta}{\gamma + 1} \left(\frac{3}{2}\right)^{2/3} \left[z_1^{2/3} + \int_0^\infty z^{2/3} \frac{\partial x}{\partial z} \, dz\right] \quad (17)$$

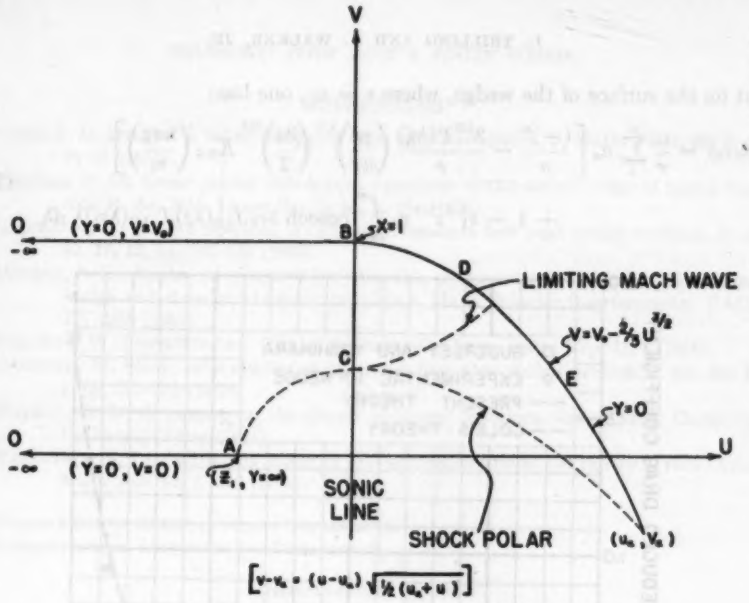


FIG. 2. Wedge problem in hodograph plane

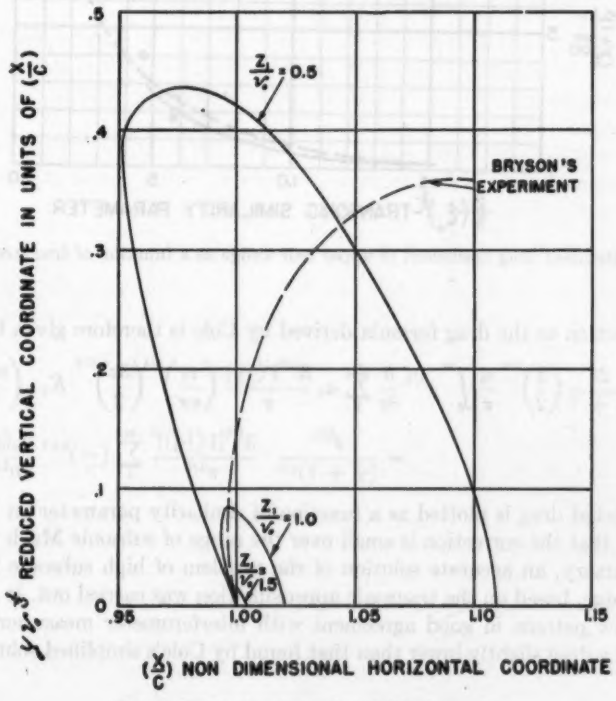


FIG. 3. Sonic sine for reduced Mach number

But on the surface of the wedge, where $v = v_0$, one has:

$$x(z_1 v_0) = \frac{v_0}{\pi} \sum_1^\infty a_n \left[\frac{(-)^n}{n} - \frac{3^{1/6}\Gamma(\frac{1}{3})}{\pi} \left(\frac{v_0}{n\pi}\right)^{1/3} \left(\frac{3z}{2}\right)^{2/3} K_{2/3}\left(\frac{n\pi z}{v_0}\right) \right] + 1 - z_1^{1/3} z^{2/3} v_0 \int_0^\infty \operatorname{cosech} \lambda v_0 J_{2/3}(\lambda z) J_{-1/3}(\lambda z_1) \lambda \, d\lambda \tag{18}$$

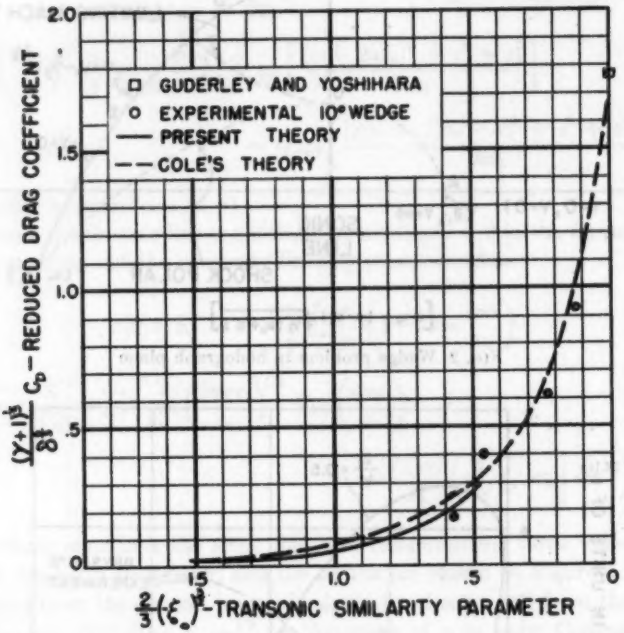


FIG. 4. Reduced drag coefficient of upper half wedge as a function of free stream Mach number.

The correction to the drag formula derived by Cole is therefore given by

$$\Delta C_D = \frac{2\delta}{\gamma+1} \left(\frac{3}{2}\right)^{2/3} \frac{v_0}{\pi} \int_0^\infty z^{2/3} \frac{\partial}{\partial z} \sum_1^\infty a_n \frac{3^{1/6}\Gamma(\frac{1}{3})}{\pi} \left(\frac{v_0}{n\pi}\right)^{1/3} \left(\frac{3z}{2}\right)^{2/3} K_{2/3}\left(\frac{n\pi z}{v_0}\right) dz$$

$$= \frac{\delta^{5/3}}{(\gamma+1)^{1/3}} \cdot \frac{3^{1/6}[\Gamma(\frac{1}{3})]^2}{\pi^{5/3}} \sum_1^\infty (-)^{n+1} \frac{a_n(\xi_0)}{n^{5/3}} \tag{19}$$

The corrected drag is plotted as a function of similarity parameter on Figure 4. It is seen that the correction is small over the range of subsonic Mach numbers.

In summary, an accurate solution of the problem of high subsonic flow past a thin wedge, based on the transonic approximation was carried out. It showed a mixed flow pattern in good agreement with interferometer measurement, and predicted a drag slightly lower than that found by Cole's simplified solution.

REFERENCES

1. Cole, J. D. Drag of a finite wedge at high subsonic speeds. J. Math. Phys. **30**, 3, pp. 79–93 (1951).
2. Tricomi, F. On linear partial differential equations of the second order of mixed type. Atti R. Ac. Noz. Lincei Ser. 5, **14**, p. 134 (1923).
3. Liepmann, H. W. and Bryson, A. E., Jr. Transonic flow past wedge sections. J. Ae. Sc. **17**, 12, pp. 745–755 (1950).
4. Bryson, A. E., Jr. An experimental investigation of transonic flow past two-dimensional wedge and circular arc sections using a Mach Zehnder interferometer NACA-TN-2560 (1951).
5. Griffith, W. Transonic flow. Princeton University Phys. Dept. Rep. II, 7 (1950).
6. Griffith, W. Shock tube studies of transonic flow over wedge profiles. J. Ae. Sc. **19**, 4, pp. 249–258 (1952).
7. Watson, G. N. A treatise on the theory of Bessel functions. 2nd Edition Cambridge University Press (1948).
8. Weinstein, A. Discontinuous integrals and generalized potential theory. Trans. Amer. Math. Soc. **63**, p. 342 (1948).

Department of Aeronautical Engineering
Massachusetts Institute of Technology

(Received June 24, 1952)

ON A STABILITY CRITERION FOR PARTIAL DIFFERENCE EQUATIONS

By Stephen H. Crandall

In a recent paper, O'Brien, Hyman and Kaplan [1] describe von Neumann's stability analysis for linear partial difference equations with constant coefficients. For non-linear equations or equations with variable coefficients they suggest testing the stability point by point in a sequence of overlapping small regions. Admitting the heuristic nature of this procedure, they claim it has worked well in practice. The purpose of the present note is to show that in general it is *not possible* to make stability deductions from successive local investigations. We give an example[1] of a linear equation with almost constant coefficients for which von Neumann's stability criterion is satisfied at every point but which has solutions which grow exponentially.

Consider the differential equation

$$(1 - 2\epsilon \cos 2t)\, \partial^2\psi/\partial x^2 = \partial^2\psi/\partial t^2 \tag{1}$$

with $0 < \epsilon \ll 1$, and the corresponding difference equation

$$r^2(1 - 2\epsilon \cos 2t)(\phi_{j+1,k} - 2\phi_{j,k} + \phi_{j-1,k}) = \phi_{j,k+1} - 2\phi_{j,k} + \phi_{j,k-1} \tag{2}$$

with $r = \Delta t/\Delta x$. Using the method described in [1] it is easy to show that (2) will satisfy their stability criterion at every point if $r^2 < 1/(1 + 2\epsilon)$. Accordingly it might then be assumed that all solutions of (2) are bounded or, at worst, grow linearly with t.

Let us consider the solutions of (1) and (2) for $t > 0$ with the initial conditions, $\psi = \cos x$ and $\partial\psi/\partial t = 0$, on $t = 0$. The values of the solution of (1) along $x = 0$ will then satisfy the ordinary differential equation

$$d^2\psi/dt^2 + (1 - 2\epsilon \cos 2t)\psi = 0 \tag{3}$$

with the initial conditions $\psi = 1$, $d\psi/dt = 0$ at $t = 0$. Similarly the values of the solution of (2) along $x = 0$ will satisfy the ordinary difference equation

$$\phi_{0,k+1} - 2\phi_{0,k} + \phi_{0,k-1} + \frac{\pi^2}{n^2}\left(\frac{\sin \pi/2nr}{\pi/2nr}\right)^2\left(1 - 2\epsilon \cos \frac{2k\pi}{n}\right)\phi_{0,k} = 0 \tag{4}$$

with the initial conditions, $\phi_{0,0} = 1$ and $\phi_{0,1} = \phi_{0,-1}$. The interval Δt has been set equal to π/n. We can show that for any $r > 0$ and n large enough, $\phi_{0,k}$ will grow exponentially. This may be done directly using the theory of difference equations with periodic coefficients [2]. A less tedious, but indirect argument relates the instability of (4) to that of (3). From the theory of Mathieu's equation [3] we know that at $t = m\pi$ the solution of (3) takes on the values $\psi = (-1)^m \cosh m\mu\pi$ where $\mu \approx \epsilon/2$. The analogous theory [2, p. 205] for (4) indicates

[1] This idea grew out of a discussion between Norman Levinson and the author in which Levinson pointed out the inadequacy of local criteria for ordinary differential equations.

that its solution will grow exponentially if $|\phi_{0,n}| > 1$. Now as $n \to \infty$ the solution of (4) approaches the solution of (3) uniformly [2, p. 164] in $0 \leq t \leq \pi$. Hence, $\phi_{0,n} \to -\cosh \mu\pi \approx -(1 + \pi^2\epsilon^2/4)$ and there exists an N such that for all $n > N$ the solutions of (4) grow exponentially with t.

REFERENCES

1. G. G. O'Brien, M. A. Hyman and S. Kaplan, A study of the numerical solution of partial differential equations, J. Math. and Phys. **29**, 223–251 (1951).
2. T. Fort, "Finite Differences," Oxford, 1948.
3. N. W. McLachlan, "Theory and Applications of Mathieu Functions," Oxford, 1947, p. 120.

Department of Mechanical Engineering
Massachusetts Institute of Technology

(Received April 16, 1952)

ON THE FLOW OF AN INCOMPRESSIBLE VISCOUS FLUID PAST A FLAT PLATE AT MODERATE REYNOLDS NUMBERS*

Y. H. Kuo

1. Introduction. The problem of two-dimensional steady flow of an incompressible viscous fluid past an obstacle is today dealt with mostly by two approximate methods. For extremely small Reynolds number, it was shown by Oseen (1) that the problem can be tackled by linearization. For extremely large Reynolds number, it can be solved by simplifying the Navier-Stokes equations by the concept of the boundary layer, first conceived by Prandtl (2). Oseen's method of solution, though linear, is usually difficult in evaluation of all the arbitrary constants involved. Moreover, owing to the fact that the form of solution, even the first approximation, is more often than not complex, successive approximations are usually not possible.

On the other hand, the boundary-layer solution is simple in form and, most important, sufficiently accurate, as was well established by experiments. It might be considered advantageous to develop the process of iteration by taking boundary layer solution as the first approximation. However, as was shown by H. L. Alden (3) in the case of semi-infinite plate, if the boundary layer solution is taken as the first approximation, then all the higher approximations become progressively more singular and, therefore, the resistance of the plate cannot be defined. For this reason, this procedure, too, fails.

The difficulties of the method adopted by Alden originates from the facts that the flow field surrounding the plate can actually be subdivided into several domains, not all of which meet the conditions required by the boundary-layer theory (cf. for instance, Carrier and Lin's discussion (4)). As the boundary-layer solution is supposed to be valid near the solid surface, in the neighborhood of the leading edge where the viscous region extends equally in the forward direction about the edge, the boundary layer approximations—namely, that the rates of change of the flow properties along and perpendicular to the plate differ by one order of magnitude—clearly break down. Poor approximation there inescapably causes a singularity to appear. If this solution is chosen as the first approximation to the dynamical equations, the higher approximations which satisfy differential equations containing highly singular terms which were absent in the first approximation, must be necessarily more singular. Thus, as the order of approximation increases, higher-order singularities unavoidably appear and, consequently, the drag integral must diverge.

This difficulty in the theory of perturbation is, of course, not new and frequently occurs in a variety of physical problems. The problem of rendering the approximate solutions uniformly valid has led M. J. Lighthill (5) to the dis-

* This study was carried out as part of a research program at the Graduate School of Aeronautical Engineering and was sponsored by the Office of Naval Research.

covery of a method which has become an effective tool in solving non-linear problems. Basically, the idea seems to be that if the solution of an approximated equation to a non-linear total or partial differential equation does not represent uniformly a physical quantity over a domain, it can be improved in the neighborhood of the singularity by merely straining slightly the arguments of the solution. When this idea is formulated mathematically, the condition of uniformity in the analytic behavior of all approximations actually provides a unique way of straining the coordinates. The mathematical structure of the method is, therefore, by itself complete. In the example treated below, the differential equation of the first approximation is non-linear but its solution is singular at the leading edge of the plate. For this case, though not included in Lighthill's treatment, it can be shown that, without modification, Lighthill's technique can still be successfully applied. From this point of view, perhaps it is not without interest.

The first part of this paper (Part I) is devoted to the problem of flow past a finite flat plate. From the character of the external potential flow an expansion parameter is deduced, which turns out to be proportional to the maximum thickness ratio of the boundary layer. If this expansion parameter is adopted, then it can be shown that, to the second approximation, the problem of solving a viscous flow remains of the boundary-layer type. Namely, from the first-order pressure calculated from the maximum displacement of the stream-lines by viscosity, along the plate and in the wake, the second-order viscous flow in the boundary layer can be determined.

The first-order potential flow outside the boundary layer is shown to be very similar, especially near the leading edge, to that given by the linearized theory of flow over a thin airfoil with round leading edge in an inviscid fluid. At the trailing edge where the flow suffers a sudden change of direction, it produces a strong expansion by the appearance of a logarithmic singularity.

The complete solution of the second-order viscous flow leads to a law of resistance which coincides with the Blasius law for Reynolds number over 10^4. In the range of Reynolds numbers from 15 to 10^4, it agrees perfectly with the measurements of Z. Janour (6).

In Part II of the present work, the main problem is the uniformization of the analytic behavior of the successive approximations, taking the boundary layer solution as the basic approximation. It is found that in the case of a flat plate, if Lighthill's conditions are imposed to the third approximation (as the second approximation needs no modification), there results a second-order shift of one coordinate x, say, measured along the plate. According to this transformation, the original constant-velocity lines, being a family of parabolas with vertices at the leading edge, become a different family of parabolas whose vertices are now all separated and located ahead of the leading edge. Furthermore, the singular line which was formerly normal to the plate, passing through the leading edge, has now been completely eliminated. The interesting fact that, except at the leading edge, the improved Blasius solution will be nowhere singular, was expected to bring out the effect of the leading edge. This was investigated and, in-

deed, was the case. The flow in the immediate neighborhood of the leading edge is found to satisfy Stokes equation of slow viscous motion, as it should.

In the light of this discussion, two features of the boundary layer approximations deserve to be mentioned. First the dimension of the viscous region around the leading edge in which the boundary layer hypotheses are violated is actually of the order of ν/U_∞ where ν is the kinematic viscosity and U_∞ the forward speed and hence is extremely small, compared with the length of the plate when the Reynolds number is large. If the detailed information, such as velocity distribution, is not required, the significance of that region, for all practical purposes, could be ignored. Secondly, as the effect of straining the coordinates vanishes on the plate, the very existence of a finite viscous region about the plate does not affect the law of resistance deduced by boundary-layer theory, even to the second order approximation as calculated in Part I. This explains why, with the effect of leading edge poorly represented, the said improved boundary layer solution still yields a correct law of resistance for Reynolds number greater than 15.

2. Navier-Stokes equations and method of expansion. The problem of steady two-dimensional flow of an incompressible viscous fluid past a fixed infinite flat plate of chord length L will be treated here by expanding the velocity field in power series of a parameter ϵ. Let the rectangular Cartesian coordinate system $(\bar{x}, \bar{y})$ be adopted and assume that the plate is defined by $0 \leqq \bar{x} \leqq L$ and $\bar{y} = 0$. Referring to this system, the $\bar{x}$- and $\bar{y}$-components of the velocity will be denoted by $\bar{u}$ and $\bar{v}$ respectively and the pressure by $\bar{p}$. If the density ρ and the kinematic viscosity ν of the fluid are constant and if the non-dimensional variables are introduced:

$$\begin{aligned} u &= \bar{u}/U_\infty, \qquad v = Re^{\frac{1}{2}}\bar{v}/U_\infty = Re^{\frac{1}{2}}V, \qquad p = (\bar{p} - \bar{p}_\infty)/\rho U_\infty^2 \\ x &= \bar{x}/L, \qquad y = Re^{\frac{1}{2}}\bar{y}/U_\infty = Re^{\frac{1}{2}}Y, \qquad Re = U_\infty L/\nu \end{aligned} \tag{2.1}$$

where U_∞ and $\bar{p}_\infty$ are, respectively, the velocity and pressure at infinity, and Re, the Reynolds number; then the velocity and pressure are determined by the system:

$$\begin{aligned} \psi_y\psi_{xy} - \psi_x\psi_{yy} - \psi_{yyy} &= -p_x + Re^{-1}\psi_{yxx} \\ p_y + Re^{-1}(\psi_x\psi_{xy} - \psi_y\psi_{xx} + \psi_{xyy}) &= -Re^{-2}\psi_{xxx} \end{aligned} \tag{2.2}$$

and

$$u = \psi_y, \qquad v = -\psi_x$$

where ψ is the non-dimensional stream function and the subscripts denote partial differentiation with respect to the variable indicated.

To solve this problem from the standpoint of perturbation, it is assumed that the stream function ψ and the pressure p can be expanded in the forms:

$$\begin{aligned} \psi &= \psi^{(0)} + \epsilon\psi^{(1)} + \epsilon^2\psi^{(2)} + \cdots \\ p &= p^{(0)} + \epsilon p^{(1)} + \epsilon^2 p^{(2)} + \cdots \end{aligned} \tag{2.3}$$

If the boundary-layer flow is taken as the first approximation to the flow in question, $\psi^{(0)}$ then satisfies

$$\psi_y^{(0)}\psi_{xy}^{(0)} - \psi_x^{(0)}\psi_{yy}^{(0)} = \psi_{yyy}^{(0)} \tag{2.4}$$

as the pressure $p^{(0)} = 0$ in the case of flat plate. As will subsequently be developed, to the second approximation, the fundamental boundary-layer hypothesis still holds; namely that the viscous effects take place inside a narrow region along the solid surface, so that the flow outside this layer is essentially frictionless and irrotational. If the two flow fields join smoothly, the vertical velocity in the potential field, which is, by (2.3), $O(\epsilon)$ must equal to that of the viscous layer which is $O(Re^{-\frac{1}{2}})$ by boundary-layer theory. Therefore, ϵ must be $O(Re^{-\frac{1}{2}})$ and, for definiteness, can be taken as $Re^{-\frac{1}{2}}$. The parameter ϵ as defined in this manner does have the significance of a thickness ratio, as will be clear later.

By using this definition of ϵ, the eqs. (2.2), after the substitution of ψ and p from (2.3), can be expanded in power series of ϵ. On equating the coefficients of ϵ^n of both series to zero, there result two equations for $\psi^{(n)}$ and $p^{(n)}$, $n = 1, 2, 3, \cdots$. It is the purpose of Part I to deal with the second approximation only, the equations for n greater than or equal to 2 will not be given. For $n = 1$, these equations read:

$$\begin{aligned} \psi_y^{(0)}\psi_{xy}^{(1)} + \psi_{xy}^{(0)}\psi_y^{(1)} - \psi_x^{(0)}\psi_{yy}^{(1)} - \psi_{yy}^{(0)}\psi_x^{(1)} &= -p_x^{(1)} + \psi_{yyy}^{(1)} \\ 0 &= -p_y^{(1)} \end{aligned} \tag{2.5}$$

From the second of Eq. (2.5), it follows immediately that $p^{(1)}$ is a function of x and hence is constant inside the boundary layer at any fixed point x and joins the external pressure at the edge of the boundary layer. In other words, to the second approximation the problem of viscous flow past a flat plate remains within the framework of boundary-layer theory.

If $U_1(x)$ denotes the first-order velocity of the potential field on the edge of boundary layer, then from Bernoulli's equation, the first-order pressure is

$$p^{(1)} = -U_1(x) \tag{2.6}$$

Thus, the elimination of $p^{(1)}$ from Eq. (2.5) yields for $\psi^{(1)}$ an equation:

$$\psi_y^{(0)}\psi_{xy}^{(1)} + \psi_{xy}^{(0)}\psi_y^{(1)} - \psi_x^{(0)}\psi_{yy}^{(1)} - \psi_{yy}^{(0)}\psi_x^{(1)} = U_1'(x) + \psi_{yyy}^{(1)} \tag{2.7}$$

where the prime indicates total differentiation. This equation can now be integrated if the outside flow field is determined. The boundary conditions for any order of approximation can be expressed as:

$$\begin{aligned} \psi^{(n)} = \psi_y^{(n)} &= 0, \qquad \text{when} \quad y = 0 \\ \psi_y^{(n)} &= U_n(x), \qquad \text{when} \quad y = \infty \end{aligned} \tag{2.8}$$

3. First-order potential solution. In the preceding section, it has been shown that to the second approximation the flow field can still be subdivided into a viscous layer and a potential region. By choosing the boundary-layer solution as the first approximation, the first-order potential flow can now be determined.

As was shown by Blasius' solution (7), at the edge of the boundary layer the velocity u approaches the value one but $\bar{v}/U_\infty$ tends asymptotically to $(\frac{1}{2})\beta\,\epsilon\,x^{-\frac{1}{2}}$, β being a pure number (1.73 approximately). The fact that the velocity on the edge of the viscous layer is variable shows that the flow outside it must be deflected and therefore cannot be uniform. This will be the basic principle upon which the first-order potential solution shall be constructed.

The problem can thus be formulated by the potential theory in terms of the boundary value of the vertical velocity on the x-axis. First of all, it is zero for $x < 0$. In the interval $0 \leqq x \leqq 1$, $V = (\frac{1}{2})\beta\,\epsilon\,x^{-\frac{1}{2}}$, according to boundary-layer theory. On account of the wake behind the plate, however, there is no fore-and-aft symmetry and its effect should be separately considered.

According to Goldstein's theoretical study and Fage's experiments on the wake behind a flat plate (8), the boundary of the wake begins with the boundary layer at the trailing edge of the plate and grows rather slowly downstream as long as it remains laminar. The vertical velocity at the boundary of the wake varies as $-(x-1)^{1/3}$(9) in the neighborhood of the plate and asymptotically as $-(x-0.48)^{-3/2}$(10) at large distances downstream. Thus, it is seen that the limiting streamlines change suddenly from positive to negative inclinations with a discontinutiy at the trailing edge of the plate. From this incomplete information about the wake, the boundary value on the rest of the x-axis can only be approximately described. This has been carried out. However, owing to the fact that V varies slowly for x close to one and the coefficient of $(x-0.48)^{-3/2}$ is small, the final result is very little affected if the wake is replaced by a discontinuity in the inclination of the stream-lines at the trailing edge and remains parallel from there on. Namely, on the x-axis:

$$V = \begin{cases} 0 & \text{for} \quad x < 0 \\ (\frac{1}{2})\beta\,\epsilon\,x^{-\frac{1}{2}} & \text{for} \quad 0 \leqq x \leqq 1 \\ 0 & \text{for} \quad 1 < x \leqq \infty \end{cases} \tag{3.1}$$

If the velocity field in the potential region can be expanded in terms of a thickness parameter ϵ like the function ψ, then

$$\begin{aligned} u &= 1 + \epsilon U^{(1)}(\bar{x}/L, \bar{y}/L) + \cdots \\ V &= \epsilon V^{(1)}(\bar{x}/L, \bar{y}/L) + \cdots . \end{aligned} \tag{3.2}$$

From the potential theory, the first-order velocities $U^{(1)}$ and $V^{(1)}$, which vanish at infinity, are given, on account of (3.1), by the following:

$$\begin{aligned} U^{(1)} - iV^{(1)} &= \frac{\beta}{2\pi}\int_0^1 \frac{dt}{t^{\frac{1}{2}}(z-t)}, \\ &= -\frac{i\beta}{2\sqrt{z}} + \frac{\beta}{2\pi\sqrt{z}}\ln\frac{1+\sqrt{z}}{1-\sqrt{z}} \end{aligned} \tag{3.3}$$

where the complex variable z denotes $(x + iY)$.

From (3.3), the velocity $U^{(1)}$ reduces, on the plate $0 \leqq x \leqq 1$, to

$$U^{(1)}(x, 0) \equiv U_1(x) = \frac{\beta}{2\pi} \frac{1}{\sqrt{x}} \ln \frac{1+\sqrt{x}}{1-\sqrt{x}}, \\ = \frac{\beta}{\pi}\left[1 + \frac{x}{3} + \frac{x^2}{5} + \cdots\right]. \tag{3.4}$$

The first-order pressure $p^{(1)}$ on the plate, following Eq. (2.6), reads:

$$p^{(1)} \equiv p_1(x) = -(\beta/\pi)[1 + x/3 + x^2/5 + \cdots .] \tag{3.5}$$

It is noted that due to the displacement of the stream-lines from the plate by the viscous effect, the resulting potential flow along the plate is accelerating. Consequently, the viscous layer, in turn, is subject to a favorable pressure gradient in the downstream direction.* Moreover, in spite of the fact that the vertical velocity $V^{(1)}$ is singular at the leading edge, the horizontal velocity $U^{(1)}$ and, hence the pressure, are regular there. On the other hand, $V^{(1)}$ is discontinuous but finite at the trailing edge, where $U^{(1)}$ has a logarithmic singularity. The consequence of these is that all the derivatives of the velocities, such as vorticity, in the viscous layer will be even more singular at both ends of the plate. This seems to be in agreement with the solutions obtained by Oseen's approximation. (11).

4. Solution of $\psi^{(1)}(x, y)$. After the velocity at the edge of the boundary layer has been determined, the equation (2.7) can be written as

$$\psi_y^{(0)}\psi_{xy}^{(1)} + \psi_{xy}^{(0)}\psi_y^{(1)} - \psi_x^{(0)}\psi_{yy}^{(1)} - \psi_{yy}^{(0)}\psi_x^{(1)} - \psi_{yyy}^{(1)} \\ = (\beta/\pi)[1/3 + 2/5\, x + \cdots + [(n-1)/(2n-1)]x^{n-2} + \cdots] \tag{4.1}$$

By hypothesis, $\psi^{(0)}$ satisfies Eq. (2.6) and, according to Blasius, $\psi^{(0)} = x^{\frac{1}{2}} f_0(\eta)$ and $\eta = y/\sqrt{x}$, where the function f_0 can be obtained by integratng the equation

$$2f_0''' + f_0 f_0'' = 0 \tag{4.2}$$

These special forms of $\psi^{(0)}$ and $U_1(x)$ demand as solutions:

$$\psi^{(1)} = \frac{\beta}{\pi}\left[x^{\frac{1}{2}} f_1(\eta) + \tfrac{1}{3}x^{3/2} f_2(\eta) + \cdots + \frac{x^{n-\frac{1}{2}}}{2n-1} f_n(\eta) + \cdots\right] \tag{4.3}$$

* Note that this pressure gradient vanishes in the limit of an infinite plate. This is because the displacement thickness contour is then a parabola, and it is well known that in the flow past a parabola, the pressure on the surface becomes constant and equal to the mainstream pressure at distances from the leading edge compared with the focal length. (Also the pressure continued onto the axis is exactly constant to the right of the focus.) In the assumed flow (resulting from the finite length of the plate), however, the parabola is broken off at the trailing edge and becomes a line parallel to the stream. This breaking off is effectively like a sink distribution in the wake, and results in the accelerated flow along the plate which was calculated above.

whence

$$\psi_y^{(1)} = \frac{\beta}{\pi}\left[f_1' + \cdots + \frac{x^{n-1}}{2n-1} f_n' + \cdots\right]$$
$$\psi_x^{(1)} = \frac{\beta}{\pi}\left[\tfrac{1}{2}x^{-\frac{1}{2}}(f_1 - \eta f_1') + \cdots + \frac{x^{n-3/2}}{2(2n-1)}\{(2n-1)f_n - \eta f_n'\} + \cdots\right] \tag{4.4}$$

where $f_n(\eta)$'s are functions of η.

The substitution of $\psi^{(1)}$ from (4.3) in (4.1), after a simple reduction, gives:

$$2f_n''' + f_0 f_n'' - 2(n-1)f_0' f_n' + (2n-1)f_0'' f_n = -2(n-1) \tag{4.5}$$

where $n = 1, 2, 3, \cdots$. From the boundary conditions (2.8) and on account of (3.4), the functions f_n satisfy, for all n and $x > 0$:

$$f_n(0) = f_n'(0) = 0, \qquad f_n'(\infty) = 1 \tag{4.6}$$

As the coefficients of this equation cannot be expressed analytically by simple functions for the whole range of η, an analytical solution valid for all η and n is generally not possible. The problem, on the whole, is a numerical one. It is noticed, however, that for $n = 1$, the Eq. (4.5) assumes a particularly simple form:

$$2f_1''' + f_0 f_1'' + f_0'' f_1 = 0 \tag{4.7}$$

For this equation, it can be verified that the following combination of f_0 and f_0', which evidently satisfies (4.6), is a solution:

$$f_1 = \tfrac{1}{2}(f_0 + \eta f_0') \tag{4.8}$$

The cases $n = 2$ and $n = 3, 5, 9$ have been integrated numerically by Howarth (12) and Tani (13) respectively. It will be shown presently that for n greater than 10, say, simple approximate solutions can be found.

First of all, it is noticed that f_0' is a special solution of the equation:

$$2f_n''' + f_0 f_n'' - 2(n-1)f_0' f_n' + (2n-1)f_0'' f_n = 0$$

Therefore, the order of the Eq. (4.5) can be reduced by one if the substitution:

$$f_n = f_0' F_n(\eta) \tag{4.9}$$

is made. Elimination of f_n from Eq. (4.5) by virtue of (4.9) leads to:

$$2f_0' F_n''' + (6f_0'' + f_0 f_0')F_n'' - (f_0 f_0'' + 2(n-1)f_0'^2)F_n' = -2(n-1) \tag{4.10}$$

This equation can then be considered as second order in F_n'. If the coefficients of this equation are to be approximated by taking $f_0 = (\frac{1}{2})\alpha\eta^2$ in a range of $\eta \leqq \eta_1$, where α is 0.332 and η_1 is a fixed number to be specified later; then in the range $0 \leqq \eta \leqq \eta_1$, Eq. (4.10) will be replaced by

$$\eta^2 F_n''' + 3\eta F_n'' - (n - \tfrac{3}{4})\alpha\eta^3 F_n' = -\frac{n-1}{\alpha}\eta \tag{4.11}$$

The errors involved in the first and third coefficients are of the order η^3 and that of the second is of the order η^5. Therefore, if η_1 is not too large, this approximation is expected to be sufficiently accurate for the answers required.

It is observed, in the first place, that $(n-1)/\alpha^2(n-\frac{3}{4})\eta^2$ is a particular solution. Furthermore, the fundamental integrals of the homogeneous equation resulting from the cancellation of the term on the right-hand side of Eq. (4.11), are $n^{-1}I_{2/3}(\zeta)$ and $\eta^{-1}K_{2/3}(\zeta)$. The general solution then can be written as

$$F'_n = A\eta^{-1}I_{2/3}(\zeta) + C_1\eta^{-1}K_{2/3}(\zeta) + \frac{n-1}{\alpha^2(n-\frac{3}{4})}\frac{1}{\eta^2} \tag{4.12}$$

where $$\zeta = \tfrac{2}{3}(n-\tfrac{3}{4})^{\frac{1}{2}}\alpha^{\frac{1}{2}}\eta^{3/2}$$

Here $I_{2/3}$ and $K_{2/3}$ denote the Bessel functions of the first and second kinds of imaginary argument. In order that the condition at $\eta = 0$ be satisfied, the constant C_1 should be determined to cancel the singularity of $K_{2/3}$ which is η^{-1}. This condition requires that

$$C_1 = -\frac{n-1}{\alpha^2(n-\frac{3}{4})} \cdot \frac{\Gamma(\frac{1}{3})(n-\frac{3}{4})^{1/3}\alpha^{1/3}}{\pi 3^{1/6}}$$

Thus, the solution, with the conditions at $\eta = 0$ satisfied, is finally of the form:

$$f_n = f'_0\left\{A\int_0^\eta I_{2/3}(\zeta)\frac{d\eta}{\eta} - \frac{n-1}{\alpha^2(n-\frac{3}{4})}\int_0^\eta\left[\lambda_n K_{2/3}(\zeta) - \frac{1}{\eta}\right]\frac{d\eta}{\eta}\right\} \tag{4.13}$$

by putting $\lambda_n = \Gamma(\frac{1}{3})(n-\frac{3}{4})^{1/3}\alpha^{1/3}/\pi 3^{1/6}$, where the constant A remains to be determined.

On the other hand, when η is large, f_0 is approximately $\eta - \beta$. If this approximation is introduced in Eq. (4.10), it is transformed into

$$2F'''_n + (\eta-\beta)F''_n - 2(n-1)F'_n = -2(n-1) \tag{4.14}$$

Moreover, it can be shown that by the substitutions:

$$F'_n = (\eta-\beta)e^{-(\eta-\beta)^2/4}G_n(\sigma) + 1 \tag{4.15}$$

and

$$\sigma = \tfrac{1}{4}(\eta-\beta)^2$$

Eq. (4.14) becomes a confluent hypergeometric equation for $G_n(\sigma)$:

$$\sigma G''_n + (\tfrac{1}{2}+1-\sigma)G'_n - nG_n = 0 \tag{4.16}$$

whose two fundamental solutions are:

$${}_1F_1(n, \tfrac{3}{2}; \sigma) \quad \text{and} \quad \sigma^{-\frac{1}{2}}{}_1F_1(n-\tfrac{1}{2}, \tfrac{1}{2}; \sigma)$$

Therefore, the general solution for large η, by (4.15), can be expressed as

$$f'_n \equiv F'_n = (\eta-\beta)e^{-\sigma}[C_2{}_1F_1(n, \tfrac{3}{2}; \sigma) + C\sigma^{-\frac{1}{2}}{}_1F_1(n-\tfrac{1}{2}, \tfrac{1}{2}; \sigma)] + 1 \tag{4.17}$$

where C_2 and C are constants.

By virtue of the asymptotic behavior of the confluent hypergeometric function, namely as $\sigma \to \infty$

$$ {}_1F_1(a, b; \sigma) \approx [\Gamma(b)/\Gamma(a)]e^{\sigma}\sigma^{a-b}\,{}_2F_0(b - a, 1 - a; \sigma^{-1}) $$

the condition at infinity, i.e., $f_n'(\infty) = 1$, is satisfied if

$$ C_2 = -2[\Gamma(n)/\Gamma(n - \tfrac{1}{2})]C $$

By eliminating C_2 in terms of C, the solution for large η then assumes the final form:

$$ f_n = \eta + B + C\int_\eta^\infty e^{-\sigma}[{}_1F_1(n - \tfrac{1}{2}, \tfrac{1}{2}; \sigma) - 2[\Gamma(n)/\Gamma(n - \tfrac{1}{2})]\sigma^{\frac{1}{2}}\,{}_1F_1(n, \tfrac{3}{2}; \sigma)]\,d\eta \tag{4.18} $$

where B and C are constants.

If the two solutions (4.13) and (4.18) for small and large η, respectively, represent, in fact, the same function $f_n(\eta)$, then f_n, f_n' and f_n'' must be joned in a region where both solutions are valid. These three conditions are clearly both necessary and sufficient for the determination of the constants A, B and C.

5. Law of resistance. It can be shown that, up to the second approximation, the local shear remains $\rho\nu(\partial\bar{u}/\partial\bar{y})$. Thus, from Eq. (2.3) and (4.4) the shear stress τ on the plate is given by

$$ \tau = \rho U_\infty^2\,\epsilon\left[x^{-\frac{1}{2}}f_0''(0) + \frac{\beta\epsilon}{\pi}\sum_1^\infty \frac{x^{n-1-\frac{1}{2}}}{2n-1}f_n''(0) + \cdots\right] \tag{5.1} $$

If the total force experienced by one side of the plate per unit width is considered, the skin-friction coefficient C_f can be defined as

$$ C_f = \frac{1}{\frac{1}{2}\rho U_\infty^2 L}\int_0^L \tau\,d\bar{x} \tag{5.2} $$

From (5.1), the integration yeilds

$$ C_f = (4\alpha/\sqrt{Re}) + (\alpha_1/Re) + \cdots \tag{5.3} $$

where

$$ \alpha_1 = \frac{4\beta}{\pi}\sum_1^\infty \frac{f_n''(0)}{(2n-1)^2} \tag{5.4} $$

The skin-friction coefficient, to the second order, consists of two terms: the first is recognized as the Blasius law and the second, which varies inversely as the Reynolds number, constitutes the correction. This shows that, as the second term is inversely proportional to the Reynolds number, it becomes extremely small by comparison when the Reynolds number is large and, consequently, the resistance follows Blasius' law. However, as the Reynolds number decreases, the correction will gradually become significant. The point at which the re-

sistance deviates distinctly from the Blasius' law, depends, of course, on the value of α_1.

To evaluate this sum, as is seen from (5.4), it requires only the initial curvature of the functions f_n. Because of the fact that the functions f_1 to f_9 are known exactly, the first nine terms can be accurately calculated. From the ninth term on, however, as the degree of accuracy in $f_n''(0)$ will be less important, the approximate values will be substituted.

Before the actual calculation of α_1 is carried out, the constant A appearing in the solution f_n should be first determined by joining the two solutions (4.13) and (4.18). It is noted that if η_1 is taken to be 3 (justified by the fact that linear approximation to f_0' differs from the exact value by only about 16 per cent) then $\zeta_1 \approx (9/2)(n - 3/4)^{\frac{1}{2}}$. Thus, only for moderate values of η will the argument of the Bessel functions in Ref. (13) be large. By means of the asymptotic form of the Bessel function, e.g. $I_{2/3}(\zeta) \approx (e^{\zeta}/\sqrt{2\pi\zeta})(1 + O(\zeta^{-1}))$ and $K_{2/3}(\zeta) \approx \sqrt{\pi}\, e^{-\zeta}/\sqrt{2\pi}(1 + O(\zeta^{-1}))$, the integrals in (4.13) can be asymptotically integrated. Furthermore, by the same asymptotic integration, namely both n and σ large, the integrand of (4.18) is equivalent asymptotically to $\exp(-\sigma/2 - \sqrt{n\sigma})$ $(1 + O(1/\sqrt{n}))$.

After these simplifications, two algebraic equations between A and C give a solution for A which can be readily shown to be of the order $n^{1/4}e^{-\zeta_1}$. For $n = 4$, say, $\zeta_1 \doteq 9$ and A will be of the order 10^{-4}. Since the first term of $f_n''(0)$ is $2\alpha A[\eta^{-1}I_{2/3}(\zeta)]_0$, which is $O(n^{7/12}e^{-\zeta_1})$, it can be neglected by comparison with the second term. Hence, for large n, $f_n''(0)$ can be approximately given by

$$f_n''(0) \sim \frac{\Gamma(1/3)}{\Gamma(2/3)} \cdot \frac{n - 1}{(3\alpha)^{1/3}(n - 3/4)^{1/3}} \tag{5.5}$$

From Howarth's and Tani's solutions, $f_2''(0)$ and $f_3''(0)$, $f_5''(0)$, $f_9''(0)$* are 2.041 and 3.177, 5.010, 7.936 respectively. The approximate formula above gives the corresponding values to be respectively 1.838 and 3.022, 4.892, 7.844. From Dr. Tani's results, $f_4''(0)$, $f_6''(0)$, $f_7''(0)$ and $f_8''(0)$ were obtained graphically with the values, 4.083, 5.800, 6.570 and 7.290 respectively.

Of course, when n increases, the accuracy will certainly be improved. This seems to be sufficient to establish the reliability of this approximation. This result, incidentally, also establishes the convergence of the sum α_1, because, for large n, $f_n''(0)$ is $O(n^{2/3})$. The sum then will converge like a Zeta function, $\zeta(1/3)$.

The actual summation of the series was performed, hence, in three separate steps: (a) the first nine terms were calculated by means of Howarth's and Tani's solutions, (b) the subsequent five terms were summed by substituting $f_n''(0)$ from (5.5) and finally (c) from the fourteenth term on, the infinite sum was converted into a definite integral by the Euler-MacLaurin formula with the integrand $(n - 1)^{2/3}/(2n - 1)^2$. Of these three steps, (b) and (c) unavoidably involve error. However, a conservative estimate places the fourth significant figure at most in doubt. The final answer is, accordingly, $\alpha_1 = 4.12$.

* Private communication.

By adopting this value in (5.3), the skin-friction coefficient C_f can be plotted as a function of Re as shown in Fig. 1. It is seen that, for Re greater than 10^4, C_f practically coincides with the Blasius' law. But in the range $10 < Re < 10^4$, they differ considerably; for instance, at $Re = 100$, the difference amounts to nearly 24 percent of the total value. In this range of moderate Reynolds numbers, the resistance of a plate has been measured by Janour (6), and, by comparison, the theoretical and experimental values agree surprisingly well provided the Reynolds number is not close to 10. Near $Re = 10$, the calculated values

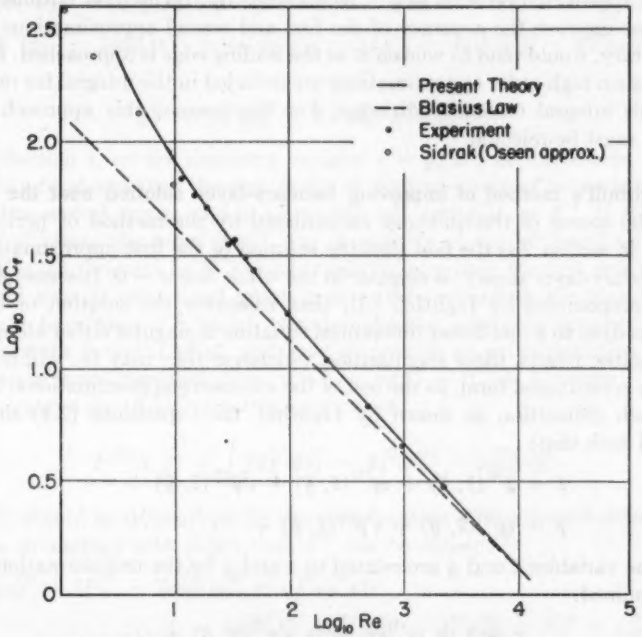


Fig. 1 Resistance coefficient of a flat plate in parallel flow.

from (5.3) are higher than the measured ones. There seems to be better agreement between the results calculated by S. Sidrak (14) by Oseen's method of approximation and the experimental values (Ref. 6) at $Re = 10$, hence it is reasonable to presume that the second-order theory cannot be accurately applied for Reynolds number below 15, say.

One might postulate that if the approximation is pursued one stage further, the range of applicability of the resistance formula (5.3) could possibly be extended into the low Reynolds number regime and might even overlap with the solution approached from the other end of the Reynolds number, such as Sidrak's solution. If this were the case, there is, at least, one example where solution can be found for all Reynolds numbers. This general question of higher approximation will be discussed in what follows.

If the approximations were carried one step further, one would notice that the known function on the right-hand side of the non-homogeneous equation for $\psi^{(2)}(x, y)$ consists of two groups of terms, having as coefficients x^{-1} and x^{-2} respectively. In view of the fact that the form of the solution has been fixed by the choice of $\psi^{(0)} = \sqrt{x} f_0(\eta)$, $\psi^{(2)}$ will automatically take on a term like $x^{-1}f(\eta)$ in order to cancel the second term on the right-hand side of the equation for $\psi^{(2)}$. Consequently, the order of singularity at the leading edge must become progressively stronger as the order of approximation increases. This shows that the high approximation such as $\psi^{(2)}$, as obtained by the method outlined above, would not improve the accuracy of the first and second approximations but, on the contrary, would tend to worsen it as the leading edge is approached. Furthermore, if such high-order approximations are included in the integral for resistance (5.2), this integral obviously diverges. For this reason, this approach to the problem must be rejected.

6. Lighthill's method of improving bundary-layer solution near the leading edge. The source of the difficulty encountered by the method of perturbation adopted in section 2 is the fact that the solution of the first approximation, i.e., the boundary-layer theory, is singular in the whole line $x = 0$. It seems generally true as propounded by Lighthill (5), that whenever the solution of the first approximation to a non-linear differential equation is singular either along curves or at isolated points, these singularities, whichever they may be, will be passed on, in an accentuated form, to the rest of the successive approximations. To overcome such difficulties, as shown by Lighthill, the expansions (2.3) should be modified such that:

$$\begin{aligned} \psi &= \psi^{(0)}(\bar{x}, \bar{y}) + \epsilon\psi^{(1)}(\bar{x}, \bar{y}) + \epsilon^2\psi^{(2)}(\bar{x}, \bar{y}) + \cdots \\ p &= \epsilon p^{(1)}(\bar{x}, \bar{y}) + \epsilon^2 p^{(2)}(\bar{x}, \bar{y}) + \cdots \end{aligned} \tag{6.1}$$

where the variables $\bar{x}$ and $\bar{y}$ are related to x and y by the transformation, yet to be determined,

$$\begin{aligned} x &= \bar{x} + \epsilon x^{(1)}(\bar{x}, \bar{y}) + \epsilon^2 x^{(2)}(\bar{x}, \bar{y}) + \cdots \\ y &= \bar{y} \end{aligned} \tag{6.2}$$

The object of the transformation (6.2) is to bring into the problem extra degrees of freedom by means of which the "parental" singularties, such as $\sqrt{x}$, can be frozen during the course of the successive approximations. This will be demonstrated in the determination of $\psi^{(2)}$ as follows:

As the solution of $\psi^{(1)}$ given in section 4 has an identical singularity to that of $\psi^{(0)}$, no further change for Eq. (2.7) is necessary. The first-order function $x^{(1)}$ of the transformation (6.2) can, therefore, be taken as zero. The transformation then becomes

$$\begin{aligned} x &= \bar{x} + \epsilon^2 x^{(2)}(\bar{x}, \bar{y}) + \cdots \\ y &= \bar{y}. \end{aligned} \tag{6.3}$$

According to this transformation, the derivatives $\partial/\partial x$ and $\partial/\partial y$ will be transformed, respectively, into $\partial/\partial\bar{x} - \epsilon^2 x_{\bar{x}}^{(2)}\,\partial/\partial\bar{x}$ and $\partial/\partial\bar{y} - \epsilon^2 x_{\bar{y}}^{(2)}\,\partial/\partial\bar{x}$. Thus, the largest terms produced by the transformation of derivatives can only be of the second order. As a consequence of this the equations for $\psi^{(0)}$ and $\psi^{(1)}$ in terms of the variables $\bar{x}$ and $\bar{y}$ will be the same as Eqs. (2.4) and (2.7). The equations for $\psi^{(2)}$ and $p^{(2)}$, however, are changed; viz.,

$$\begin{aligned}\psi_{\bar{y}}^{(0)}\psi_{\bar{x}\bar{y}}^{(2)} + \psi_{\bar{x}\bar{y}}^{(0)}\psi_{\bar{y}}^{(2)} - \psi_{\bar{x}}^{(0)}\psi_{\bar{y}\bar{y}}^{(2)} - \psi_{\bar{y}\bar{y}}^{(0)}\psi_{\bar{x}}^{(2)} - \psi_{\bar{y}\bar{y}\bar{y}}^{(2)}\\ = [\psi_{\bar{x}}^{(1)}\psi_{\bar{y}\bar{y}}^{(1)} - \psi_{\bar{y}}^{(1)}\psi_{\bar{x}\bar{y}}^{(1)}] + [-p_{\bar{x}}^{(2)} + \psi_{\bar{y}\bar{x}\bar{x}}^{(0)} + \psi_{\bar{y}\bar{y}\bar{y}}^{(0)}x_{\bar{x}}^{(2)}\\ + (p_{\bar{y}}^{(2)} - 2\psi_{\bar{x}\bar{y}\bar{y}}^{(0)})x_{\bar{y}}^{(2)} + \psi_{\bar{y}}^{(0)}\psi_{\bar{x}}^{(0)}x_{\bar{x}\bar{y}}^{(2)} - (\psi_{\bar{x}}^{(0)2} + 3\psi_{\bar{x}\bar{y}}^{(0)})x_{\bar{y}\bar{y}}^{(2)} - \psi_{\bar{x}}^{(0)}x_{\bar{y}\bar{y}\bar{y}}^{(2)}]\end{aligned} \tag{6.4}$$

and

$$p_{\bar{y}}^{(2)} + \psi_{\bar{x}}^{(0)}\psi_{\bar{x}\bar{y}}^{(0)} - \psi_{\bar{y}}^{(0)}\psi_{\bar{x}\bar{x}}^{(0)} + \psi_{\bar{x}\bar{y}\bar{y}}^{(0)} = 0$$

As in Section 4, let the similarity variable $\bar{\eta} = \bar{y}/\sqrt{\bar{x}}$ be introduced. Then it can readily be shown that the singularity of the first group of terms is $\bar{x}^{-1}$ while that of the second group, due to terms like $p_{\bar{x}}^{(2)}$ and $\psi_{\bar{y}\bar{x}\bar{x}}^{(0)}$, is $\bar{x}^{-2}$. If the singularity of $\psi^{(2)}$ were the same as $\psi^{(0)}$, the singularity of the left-hand side of the first of Eqs. (6.4) would be $\bar{x}^{-1}$, which could balance only the first group on the right-hand side. If the second group is also to be cancelled, the singularity of $\psi^{(2)}$ should have been $\bar{x}^{-\frac{1}{2}}$, which would make the resistance integral diverge. The role of the function $x^{(2)}(\bar{x}, \bar{y})$ is just to provide an extra condition to have this group of terms removed. Now, if $p^{(2)} = p_2(\bar{x}) + P^{(2)}(\bar{x}, \bar{y})$ where $P^{(2)}(\bar{x}, \bar{y})$, from the second eq. (6.4), by partial integration

$$P^{(2)}(\bar{x}, \bar{y}) = \int (\psi_{\bar{y}}^{(0)}\psi_{\bar{x}\bar{x}}^{(0)} - \psi_{\bar{x}}^{(0)}\psi_{\bar{x}\bar{y}}^{(0)} - \psi_{\bar{x}\bar{y}\bar{y}}^{(0)})\,d\bar{y} \tag{6.5}$$

and $p_2(\bar{x})$ should be determined by the consideration of the second-order potential flow, in analogy with $p_1(\bar{x})$; then $x^{(2)}$ can be chosen:

$$\begin{aligned}\psi_{\bar{y}\bar{y}\bar{y}}^{(0)}x_{\bar{x}}^{(2)} + (P_{\bar{y}}^{(2)} - 2\psi_{\bar{x}\bar{y}\bar{y}}^{(0)})x_{\bar{y}}^{(2)} + \psi_{\bar{y}}^{(0)}\psi_{\bar{x}}^{(0)}x_{\bar{x}\bar{y}}^{(2)}\\ - (\psi_{\bar{x}}^{(0)2} + 3\psi_{\bar{x}\bar{y}}^{(0)})x_{\bar{y}\bar{y}}^{(2)} - \psi_{\bar{x}}^{(0)}x_{\bar{y}\bar{y}\bar{y}}^{(2)} = P_{\bar{x}}^{(2)} - \psi_{\bar{x}\bar{x}\bar{y}}^{(0)}\end{aligned} \tag{6.6}$$

It is noted that as $x^{(2)}$ is required to be regular at $\bar{y} = 0$, both sides of eq. (6.6) vanish with $\bar{y}$. For this reason, the constant of integration will be absorbed in the yet undetermined function $p_2(\bar{x})$ which is also required to be regular.

7. Determination of the transformation function $x^{(2)}$. By the assumption $\psi^{(0)} = \sqrt{\bar{x}}\,f_0(\bar{\eta})$, together with Eq. (6.5), a straightforward reduction leads to an equation:

$$\begin{aligned}2(f_0 - \bar{\eta}f_0')g_2''' + (f_0^2 - \bar{\eta}f_0f_0' - 6\bar{\eta}f_0'')g_2'' + [2f_0'(f_0 - \bar{\eta}f_0') - 6(f_0'' + \tfrac{1}{3}\bar{\eta}f_0''')]g_2'\\ = 6\bar{\eta}f_0'' + \bar{\eta}^2f_0''' + \tfrac{1}{2}\bar{\eta}f_0f_0' - f_0^2 + \tfrac{1}{2}\bar{\eta}^2f_0''^2\end{aligned} \tag{7.1}$$

where g_2 is a function of $\bar{\eta}$ and is defined by

$$x^{(2)} = g_2(\bar{\eta}) \tag{7.2}$$

It is noted that this equation has an integrating factor $f_0(\bar{\eta})$ and, on multiplying both sides by $f_0(\bar{\eta})$, an integration yields

$$2f_0(f_0 - \bar{\eta}f_0')g_2'' - [2f_0'(f_0 - \bar{\eta}f_0') + 4\bar{\eta}f_0f_0'' - f_0^3 + \bar{\eta}f_0^2f_0']g_2' = G(\bar{\eta}) + \text{const.} \tag{7.3}$$

where

$$G(\bar{\eta}) = \int_0^{\bar{\eta}} f_0[6\bar{\eta}f_0'' + \bar{\eta}^2f_0''' + \tfrac{1}{2}\bar{\eta}f_0f_0' - f_0^2 + \tfrac{1}{2}\bar{\eta}^2f_0'^2]\,d\bar{\eta}$$

If $g_2(\bar{\eta})$ is regular at the origin and $g_2(0) = 0$ (i.e., $\bar{y} = 0$ corresponsd to $x = \tilde{x}$ and therefore the plate remains fixed), then a unique solution is found by quadrature to be of the form:

$$g_2 = -\tfrac{1}{2}\int_0^{\bar{\eta}} \left\{ \frac{f_0}{(f_0 - \bar{\eta}f_0')^2} \exp\left[-\tfrac{1}{2}\int_0^{\bar{\eta}} f_0\,d\bar{\eta}^2\right] \int_0^{\bar{\eta}} \frac{(\bar{\eta}f_0' - f_0)}{f_0^2} \cdot \exp\left[\tfrac{1}{2}\int_0^{\bar{\eta}} f_0\,d\bar{\eta}\right] G(\bar{\eta})\,d\bar{\eta} \right\} d\bar{\eta} \tag{7.4}$$

As $G(\bar{\eta})$ is $O(\bar{\eta}^4)$ in the neighborhood of $\bar{\eta} = 0$, it follows from Eq. (7.4) that $-g_2(\bar{\eta})$ is $O(\bar{\eta}^2)$. Therefore, the imposed conditions are satisfied.

The form of the solution given by (7.4), though closed, is not, however, convenient for numerical purposes. For this reason, a power-series solution will be given below by assuming f_0 to be

$$f_0 = \tfrac{1}{2}\alpha\bar{\eta}^2 - \frac{\alpha^2}{2\cdot 5!}\bar{\eta}^5 + \frac{11\alpha^3}{4\cdot 8!}\bar{\eta}^8 - \frac{375\alpha^4}{8\cdot 11!}\bar{\eta}^{11} + \cdots$$

First, the substitution of this expression under the integral sign gives the function $G(\bar{\eta})$:

$$G = \frac{3\alpha^2}{4}\bar{\eta}^4 - \frac{3\alpha^3}{140}\bar{\eta}^7 + \frac{999\alpha^4}{20\cdot 8!}\bar{\eta}^{10} - \cdots$$

Then by assuming a proper power series for g_2, and re-arranging terms on the left-hand side of Eq. (7.3), the identification of coefficients, suppressing the arbitrary constant, yields a solution

$$g_2 = -\left[\frac{1}{2\cdot 2!}\bar{\eta}^2 - \frac{\alpha}{14\cdot 5!}\bar{\eta}^5 + \frac{7\alpha^2}{30\cdot 8!}\bar{\eta}^8 - \cdots\right] \tag{7.5}$$

valid for $\bar{\eta} \leqq \bar{\eta}_1$, $\bar{\eta}_1$, being a fixed number.

On the other hand, for large $\bar{\eta}$, f_0 is again approximated by $\bar{\eta} - \beta$. The eq. (7.1) then reduces to

$$g_2''' + \tfrac{1}{2}(\bar{\eta} - \beta)g_2'' + g_2' = -\tfrac{1}{2}(\tfrac{3}{2}\bar{\eta} - \beta) \tag{7.6}$$

By the substitutions:

$$g_2' = -\tfrac{1}{2}(\bar{\eta} - \tfrac{1}{2}\beta) + g'(t) \quad \text{and} \quad t = (1/\sqrt{2})(\bar{\eta} - \beta) \tag{7.7}$$

Eq. (7.6) can be transformed into

$$g''' + tg'' + 2g' = 0$$

This equation can be twice integrated with the result

$$g' + tg = C_1 t + C_2$$

Further integration together with (7.7) yields the solution:

$$g_2 = -\tfrac{1}{4}(\bar{\eta} - \tfrac{1}{2}\beta)^2 + C_1 + C_2 e^{-\frac{1}{4}(\bar{\eta}-\beta)^2} + C_3 e^{-(\bar{\eta}-\beta)^2/4} \int_{\bar{\eta}_1}^{(\bar{\eta}-\beta)/\sqrt{2}} e^{t^2/4}\, dt \quad (7.8)$$

If the solutions (7.5) and (7.8) are joined at $\bar{\eta} = \bar{\eta}_1 = 3$, an approximate determination gives C_1, C_2 and C_3 the values -1.901, 1.264 and 0.431 respectively. Furthermore, it is noted that

$$e^{-t^2} \int_{t_1}^{t} e^{t^2}\, dt$$

approaches zero as t becomes infinite, the function g_2 tends to negative infinity as $-\frac{1}{4}(\bar{\eta} - \beta/2)^2$. g_2 so calculated is shown in Fig. 2.

Finally, it must be remarked once more that, by introducing the transformation function $x^{(2)}$ as given above, the singularity of $\psi^{(2)}$, if the solution of the equation for $\psi^{(2)}$ were to be found, could be maintained at $\sqrt{\bar{x}}$ and consequently the third-order correction to the resistance could be calculated. From a general standpoint, the problem can thus be considered as successfully solved.

8. Properties of the transformation. As was determined in the previous sections, the transformation (6.3) can now be discussed. It will be recalled first of all that the flow field of the Blasius solution is confined to the first quadrant with the plate coinciding with the positive $\bar{x}$-axis., both variables $\bar{x}$ and $\bar{y}$ being positive. From the definition of $\bar{\eta}$, if $\bar{x} \neq 0$, $\bar{y} = 0$ corresponds to $\bar{\eta} = 0$ and it follows that $x = \bar{x}$ as $g_2(0) = 0$. That is, the positive $\bar{x}$-axis is transformed into itself. On the other hand, if $\bar{x} = 0$ but $\bar{y} > 0$, $x = -\infty$ by eq. (7.4), but when $\bar{x}$ and $\bar{y}$ vanish simultaneously, $\bar{\eta}$ becomes arbitrary. If $\bar{\eta}$ varies from zero to infinity, the whole negative x-axis will be swept by the equation $x = \epsilon^2 g_2(\bar{\eta})$. Thus, the origin in the $\bar{x}$, $\bar{y}$-plane is turned into the whole negative x-axis. For values of $\bar{x}$ different from zero, it can readily be shown that every $\bar{x}$-const. line is mapped into a curve in the x, y-plane, which begins at a point on the positive x-axis and tends to negative infinity when $\bar{y}$ increases indefinitely. Consequently, the Blasius domain is transformed into the whole upper-half x, y-plane.

According to the Blasius solution, the first order velocity is a function of $\bar{\eta}$ and, following the definition of $\bar{\eta}$, its constant values lie on a family of parabolas with vertices at the origin. Since one of these parabolas defines the boundary layer, it would be of interest to determine this curve in the x, y-plane. By eliminating $\bar{x}$ from $\bar{y} = \bar{\eta}\sqrt{\bar{x}}$ by means of (6.3), $\bar{\eta} =$ const. then corresponds to

$$y = \bar{\eta}\sqrt{x - \epsilon^2 g_2(\bar{\eta})} \quad (8.1)$$

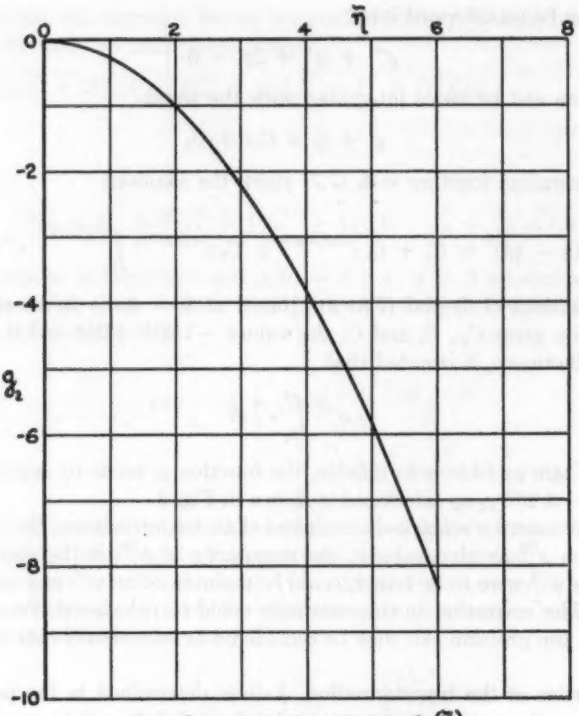


FIG. 2 THE FUNCTION $g_2(\bar{\eta})$

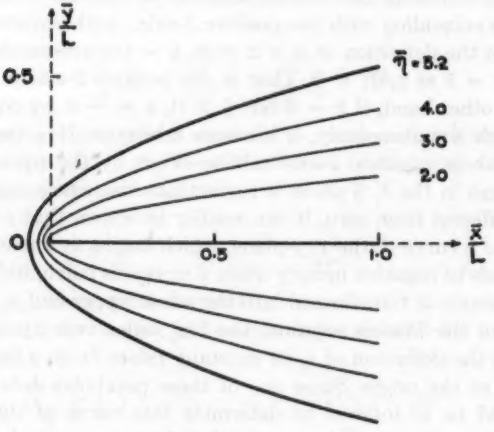


FIG. 3 BOUNDARY LAYER ON A FLAT PLATE AT R_L=100

in the x, y-planes. This shows that the family of parabolas in the $\bar{x}$, $\bar{y}$-plane remains a family of parabolas after transformation. The difference lies in the fact that before the transformation to x, y-plane the vertices of the parabolas were all located at the origin; after transformation they are separated and given by $x_0 = \epsilon^2 g_2(\bar{\eta})$ for each $\bar{\eta}$. Since $g_2(0) = 0$ and $g_2(\infty) = -\infty$, this family of parabolas will cover the whole x, y-plane. The particular parabola which characterizes the boundary layer is conventionally taken to be $\bar{\eta} = 5.2$. For this value of $\bar{\eta}$, $g_2(\bar{\eta}) = -6.30$. In the case of Reynolds number of 100, the graph is shown in Fig. 3. It is interesting to note that the region in the neighborhood of the leading edge, which was neglected in the Blasius solution, is an area of the order of ϵ^4 (linear dimension of the order of ν/U_∞). This result shows that the influence of the plate does spread ahead of the leading edge by diffusion at all Reynolds numbers and, therefore, is closer to physical reality that than predicated by the boundary-layer theory.

9. Deduced uniformly valid velocity field. A clear idea of the effects rendered to the solution by the transformation (6.3) can best be demonstrated through the examination of the velocity components in the x, y-plane. For this purpose, only the first order Blasius solution will be considered. By assuming $\psi = \psi^{(0)}$ and by making use of (6.3), the velocity components in the x, y-plane are as follows:

$$u = f_0'(\bar{\eta}) + \frac{\epsilon^2 g_2'}{2\bar{x} - \epsilon^2 \bar{\eta} g_2'} (\bar{\eta} f_0' - f_0) \tag{9.1}$$

$$V = \frac{\epsilon\sqrt{\bar{x}}}{2\bar{x} - \epsilon^2 \bar{\eta}^2 g_2'} (\bar{\eta} f_0' - f_0)$$

Since the denominator $2\bar{x} - \epsilon^2 \bar{\eta} g_2'$ in the expressions for u and V vanishes only at the leading edge, the solutions might still be singular there. However, the infinity in V on the $\bar{y}$-axis is clearly removed. Indeed, V is now finite everywhere and on the negative x-axis it is zero. The effect on u is represented by an extra term which in most of the boundary layer is of the second order. The essential change lies in the fact that, whereas previously u was unity on the negative x-axis (or on the $\bar{y}$-axis), it now varies with $\bar{\eta}$ according to (9.1) for $\bar{x} = 0$.

In order to bring out the exact nature of singularity at the leading edge, the function g_2 for small $\bar{\eta}$ will be explicitly approximated by the leading term, namely by $-\bar{\eta}^2/4$ according to (7.5). With this form of g_2, Eq. (6.3) yields a solution

$$2\bar{x} = r + x, \qquad r^2 = x^2 + Y^2 \tag{9.2}$$

on elimination of $\bar{\eta}$ by $y = \bar{\eta}\sqrt{\bar{x}}$. On account of (9.2) the same equation gives also

$$\epsilon^2 \bar{\eta}^2 = 2(r - x) \tag{9.3}$$

If the Blasius solution is approximated also by the leading term, then from (9.2) and (9.3) the velocity components valid in the neighborhood of the plate are:

$$u = \frac{\sqrt{2}\alpha}{\epsilon}\sqrt{r-x} - \frac{\alpha}{2^{3/2}\epsilon}\frac{(r-x)^{3/2}}{r}$$
$$V = \frac{\alpha}{2^{3/2}\epsilon}\sqrt{r-x}\,\frac{Y}{r} \tag{9.4}$$

From this result, it can be deduced that the stream function is proportional to the real part of $z^{3/2} - \bar{z}^{\frac{1}{2}}z$ and, therefore, is biharmonic. This shows that in the neighborhood of the leading edge, the improved solution satisfies the Stokes equation and has a singularity similar to that predicated by Lin and Carrier (4). It is noted further that the velocities u and V, which were of different order of magnitude in boundary-layer theory, are now of the same order. This is another feature of the improved Blasius solution.

By the same method, the behavior of the solution at the edge of boundary layer can be investigated. For large $\bar{\eta}$, according to (7.8), $g_2(\bar{\eta})$ can be taken to be $-\bar{\eta}^2/4$. Then it can be shown similarly that for large $\bar{\eta}$

$$u = f_0' - \frac{\epsilon\beta}{2^{3/2}}\frac{\sqrt{r-x}}{r}$$
$$V = +\frac{\epsilon\beta}{2^{3/2}}\frac{\sqrt{r+x}}{r} \tag{9.5}$$

At the edge of the boundary layer, $Y = 0(\epsilon)$. In conformity to the convention adopted in Section 3, the edge of the boundary layer will be defined by $\bar{\eta} = \infty$ or $Y = 0$. Consequently, on the boundary $Y = 0$, the velocity components will take, as boundary values, 1 and $\epsilon\beta/2\sqrt{x}$, respectively, which agree with the Blasius solution. It should be pointed out that as the edge of the boundary layer is approached, the additional term in u and V become harmonic functions, and can be represented by $-i\epsilon\beta z^{-\frac{1}{2}}$. The fact that the boundary layer solution actually tends to the character of potential flow is, indeed, remarkable. It might be added that the same conclusions apply also to the next approximations.

10. Discussion. As the results given in the preceding sections all point in the right direction, there seems no doubt that the proposed scheme is a way of solving this type of problem. The only question that one may raise is whether or not the solution so obtained is unique? For the function $g_2(\bar{\eta})$ may involve an arbitrary constant. In section 7, it was chosen to be zero by the condition that the plate remains untransformed. If the constant were not zero, the parabolas would either cut the plate or have vertices ahead of the plate (according as whether it is positive or negative). Neither case would conform to the physical picture. Therefore, the condition $g_2(0) = 0$ must be the unique choice.

In section 9, there was a result that the velocity components in the neighborhood of the plate become the same order of magnitude. From this, one might expect an increase of wall shear stress because both $\partial u/\partial y$ and $\partial v/\partial x$ have to be

considered. However, it is easy to verify that all the extra terms give no contribution at the wall. As a result the wall shear agrees with that calculated by the boundary-layer theory. This is contrary to the belief that the flow near the leading edge is of the character of slow motion, and consequently, should have higher shear stress than that calculated in the boundary-layer theory. In the light of the new results, it can be concluded that the first order solution, even after the improvement by Lighthill's method, does not give additional shear stress. The second order contribution comes solely from the fact that the plate has a finite length in the flow direction. This seems to be consistent with the consideration by momentum theorem applied to a section of the wake behind the plate.

Finally, as pointed out in Section 9, the value of the second approximation to V at the edge of the boundary layer can be defined. With this as boundary value, the second-order pressure $p^{(2)}$ can similarly be determined and the problem of solving $\psi^{(2)}$ can be, at least in principle, carried out.

REFERENCES

1. Oseen, C. W., Über die Stokessche Formel und über eine verwandte Aufgabe in der Hydrodynamik, Arch. f. mate. astro. och fysik, **6**, (1910).
2. Prandtl, L., Über Flüssigkeitsbewegung bei sehr kleiner Reibung, Proc. Third Inter. Cong. for Math. (1904) pp. 484–491.
3. Alden, Henry Leonard, Second approximation to the laminar boundary layer over a flat plate, Jour. Math. and Phys. **27**, 91–104, (1948).
4. Carrier, G. F. and Lin, C. C., On the nature of the boundary layer near the leading edge of a flat plate, Quart. Appl. Math. **6**, 63–68, (1948).
5. Lighthill, M. J., A technique for rendering approximate solutions to physical problems uniformly valid, Phil. Mag. **40**, 1179–1201, (1949).
6. Janour, Zbynek, Resistance of a plate in parallel flow at low Reynolds numbers, TM 1316 (1951).
7. Goldstein, S., Modern Developments in Fluid Dynamics, Oxford (1938) pp. 135–136.
8. Goldstein, S. (with an Appendix by A. Fage), On the two-dimensional steady flow of a viscous fluid behind a solid body. I. Proc. Roy. Soc. **A142**, 545–562, (1933).
9. Goldstein, S., Concerning some solutions of the boundary layer equations in hydrodynamics, Proc. Camb. Phil. Soc. **26**, 1–30, (1930).
10. Goldstein, S., On the two-dimensional steady flow of a viscous fluid behind a solid body. II. Proc. Roy. Soc. **A142**, 563–573, (1933).
11. Piercy, N. A. V. and Winny, H. F., The skin friction of a flat plate to Oseen's approximation, Proc. Roy. Soc. **A140**, 543–561, (1933).
12. Howarth, L., On the solution of the laminar boundary layer equations, Proc. Roy. Soc. **A164**, 547–579, (1938).
13. Tani, Itiro, On the solution of the laminar boundary layer equations, Jour. Phys. Soc. Japan **4**, 149–154, (1949).
14. Sidrak, Sobphly, The flow of a viscous liquid past an elliptic cylinder, Proc. Roy. Irish Acad. **A53**, 65–81, (1950).

Graduate School of Aeronautical Engineering
Cornell University, Ithaca, N. Y.

(Received May 4, 1953)

DIFFRACTION AND REFLECTION OF WEAK SHOCKS BY STRUCTURES

By Lu Ting

1. Introduction. The problem of the reflection of a weak shock by a flat corner was considered by Bargmann [1]. In the first approximation, in which the flow is irrotational, he obtained the velocity potential by employing the Busemann conical transformation. In a similar manner, Lighthill obtained the pressure field behind a shock (weak or strong) passing over a flat corner or a nearly vertical corner [2, 3]. Since the governing differential equations and the boundary conditions are linear, one can obtain the solution for a finite shock hitting a flat surface or a nearly vertical wall of arbitrary shape by superposition of Lighthill's solution. In a different approach, the pressure field behind a finite shock hitting a flat surface was expressed in terms of Possio integrals [4].

As another extension to Bargmann's problem, the solution for the flow behind a weak shock passing over a corner of any angle can again be obtained by using the Busemann conical transformation [5]. However, one cannot solve for the problem of a weak shock hitting an arbitrary surface by linear superposition of such solutions because the boundaries are nonlinear.

In this paper the problem of a weak shock hitting a two dimensional rectangular barrier (or structure) of height h and width L (Figs. 1 & 3) is studied from a new approach. Instead of finding a superposed solution for all times, we propose to look for solutions that satisfy all boundary conditions at suitably divided time intervals. This method is developed for structures of two different forms; that of a very thin barrier and that of a rectangular barrier of finite width (**2 & 3**).

For a weak shock, the shock strength ϵ, defined as $(P_1 - P_0)/P_0$, can be used as the expansion parameter where P_0 and P_1 are the undisturbed pressures in front of and behind the shock respectively. The disturbed pressure behind the shock is $P_0(1 + \epsilon p) + 0(\epsilon^2)$ and the velocity of the shock relative to the structure, or the air in front of the shock, is $C_0 + 0(\epsilon)$ where C_0 is the sonic velocity of the air in front of the shock.

In the first approximation, the flow is irrotational, and the velocity potential as well as the disturbance pressure, p, obeys the simple wave equation.

$$\frac{\partial^2 p}{\partial x^2} + \frac{\partial^2 p}{\partial y^2} - \frac{1}{C_0^2}\frac{\partial^2 p}{\partial t^2} = 0 \tag{1}$$

On the rigid walls the normal disturbance velocity v_n must vanish. This leads to the boundary condition $\partial p/\partial n = 0$, due to the linearized equation of motion $P_0 \partial p/\partial n = -\rho_0 \partial v_n/\partial t$, where $\rho_0 (= \gamma P_0/C_0^2)$ is the density of the air in front of the shock. The initial conditions are $p = p_t = 0$ when $t \leqq 0$.

We shall at first investigate the limiting case of $L \ll h$, i.e., the structure is in the form of a very thin barrier.

2. Thin Barriers ($L \ll h$). For the linearized theory, the thin barrier can be represented by a segment OA on the y-axis with the ground as the x-axis (Fig. 1). After hitting the barrier, $t \geqq 0$, part of the shock BD advances into still air undisturbed, while the other part EF is reflected. The pressure behind the shock EF is $P_0(1 + 2\epsilon)$ [6]. The disturbance due to the edge A is confined inside the Mach cone,

$$s_1(x, y, t) = r_1 - C_0 t = \sqrt{x^2 + (y - h)^2} - C_0 t = 0,$$

and there is no discontinuity in pressure across the Mach cone. In the following we shall find the solution for the first three time intervals

$$(0 \text{ to } h/C_0\,,\ h/C_0 \text{ to } 2h/C_0 \text{ and } 2h/C_0 \text{ to } 3h/C_0)$$

as three different problems.

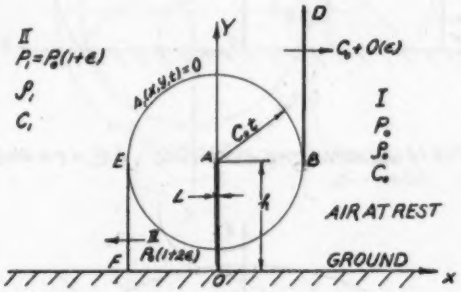


FIG. 1a DURING THE INTERVAL $0<t<h/C_0$

(a) *The First Problem.* Fig. 1a represents the shock pattern during the interval $0 \leqq t \leqq h/C_0$. The disturbance pressure will be the same, if we consider the barrier as an infinite wedge of zero wedge angle, because the removal of the ground (FOX) would not change the shock pattern.

Following the method used by Keller and Blank [5], the Busemann conical transformation $1/\zeta = \frac{1}{2}(z + z^{-1})$ with $z = Re^{i\theta}$, $\zeta = (r_1/C_0t)e^{i\theta}$, $\theta = \arctan \overline{(y - h/x)}$, and $r_1/C_0t = 2R/(1 + R^2)$, is introduced, so that the simple wave equation for p reduces to a Laplace eq.

$$p_{RR} + (1/R)p_R + (1/R^2)p_{\theta\theta} = 0$$

The boundary conditions are (Fig. 2a)

$$p = 0 \quad \text{for} \quad R = 1 \qquad 0 > \theta > -\tfrac{1}{2}\pi$$

$$p = 1 \quad \text{for} \quad R = 1 \qquad \pi > \theta > 0$$

$$p = 2 \quad \text{for} \quad R = 1 \qquad \tfrac{3}{2}\pi > \theta > \pi$$

and $$p_x = 0 \quad \text{for} \quad 0 \leqq R \leqq 1, \qquad \theta = -\tfrac{1}{2}\pi \text{ or } \tfrac{3}{2}\pi$$

In order to solve for $p(R, \theta)$ we map the exterior of the wedge in the z-plane into the upper half of the w-plane by the transformation

$$w = \Omega e^{i\omega} = (iz)^{\frac{1}{2}}.$$

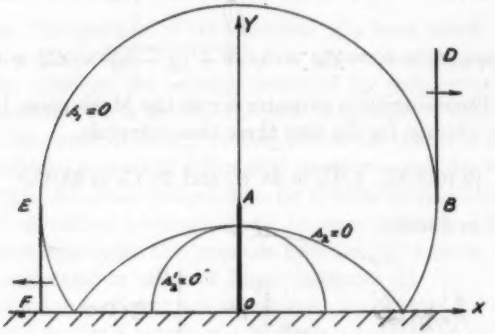


FIG. 1b DURING THE INTERVAL $h/c_0 < t < 2h/c_0$

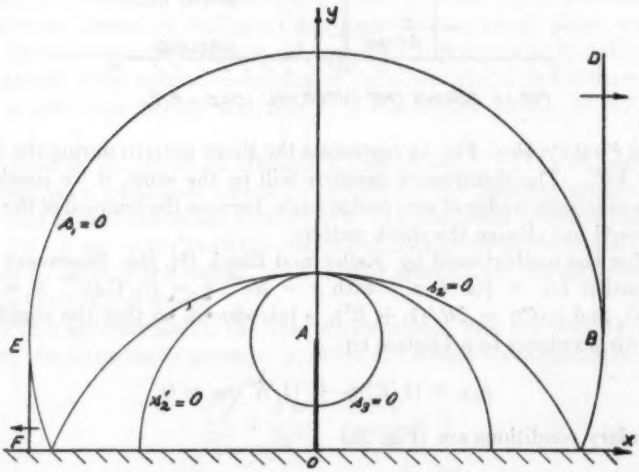


FIG. 1c DURING THE INTERVAL $2h/c_0 < t < 3h/c_0$

By the reflection principle we may extend p into the whole circle and obtain a boundary value problem in the unit circle (Fig. 2b) and the solution can be

shown to be [5]:

$$p(w) = 1 + \frac{1}{\pi}\arctan\frac{1-\Omega^2}{1+\Omega^2+2\sqrt{2}\Omega\cos\omega} - \frac{1}{\pi}\arctan\frac{1-\Omega^2}{1+\Omega^2-2\sqrt{2}\Omega\cos\omega} \tag{2}$$

The arctan is taken in the interval 0 and π. It is clear that from eq. (2) the pressure in front of and behind the barrier are different.

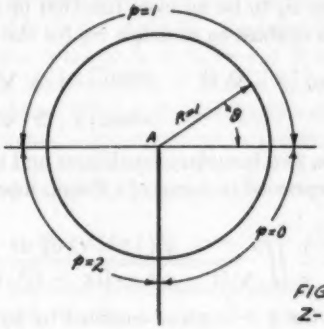


FIG. 2a
Z-PLANE

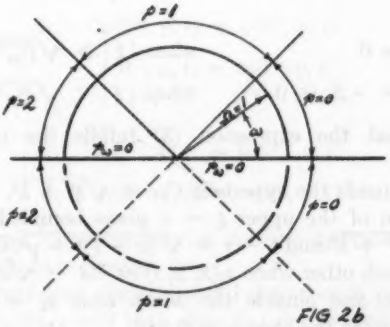


FIG 2b
W-PLANE

Eq. (2) defines p on and inside the Mach cone $s_1 = 0$. Outside the Mach cone p assumes the constant values of 0, 1, 2 in the regions I, II, III respectively (Fig. 1a). The entire solution thus obtained is continuous across the Mach cone and will be denoted by $S_1(x, y, t)$.

(b) *The Second Problem.* When $t \geqq h/C_0$, the Mach cone $s_1(x, y, t) = 0$ is intercepted and reflected by the ground (x-axis). The solution S_1 does not fulfill the boundary condition $p_y = 0$ on the x-axis. The additional disturbance S_2

in front of the structure ($x > 0$), due to the presence of the ground, has to fulfill the simple wave equation, eq. (1), and the following boundary conditions,

$$(S_2)_x = 0 \qquad \text{on the } y\text{-axis}$$

$$(S_2)_y = -(S_1)_y \qquad \text{on the } x\text{-axis, when } 0 < x < \sqrt{\overline{Ct}^2 - h^2}$$

$$(S_2)_y = 0 \qquad \text{on the } x\text{-axis, when } x > \sqrt{\overline{C_0 t}^2 - h^2}$$

and $$S_2 \to 0 \qquad \text{as } \sqrt{x^2 + y^2} \to \infty$$

Since the boundary conditions for S_2 are prescribed only on the right half of $x - t$ plane, we may choose S_2 to be an even function or x extending through the entire $x - t$ plane. This enables us to define S_{2y} for the entire $x - t$ plane:

$$S_{2y}(x, 0, t) = -S_{1y}(|x|, 0, t) \qquad \text{when } |x| < \sqrt{C_0^2 t^2 - h^2}$$

and $$S_{2y}(x, 0, t) = 0 \qquad \text{when } |x| > \sqrt{C_0^2 t^2 - h^2}$$

The solution satisfying these two boundary conditions and the condition $S_2 \to 0$ as $\sqrt{x^2 + y^2} \to \infty$, can be expressed in terms of a Possio integral (see for example ref. [9])

$$S_2(x, y, t) = -\frac{1}{\pi} \iint \frac{g_1(|\xi|, \tau)\, d\xi\, d\tau}{\sqrt{(t - \tau)^2 - [(x - \xi)^2 + y^2]/C_0^2}} \tag{3}$$

integrated over the area in the $\xi - \tau$ plane confined by hyperbola

$$t - \tau = \sqrt{(x - \xi)^2 + y^2}/C_0$$

and the ξ-axis, with

$$g_1(\xi, \tau) = 0 \qquad \text{when } |\xi| > \sqrt{\overline{C_0 \tau}^2 - h^2}$$

$$g_1(\xi, \tau) = -S_{1y}(\xi, 0, \tau) \qquad \text{when } |\xi| < \sqrt{\overline{C_0 \tau}^2 - h^2}$$

It can be shown that the expression (3) fulfills the boundary condition $S_{2x}(0, y, t) = 0$.

Since $g_1(\xi, \tau) = 0$ outside the hyperbola $C_0\tau = \sqrt{\xi^2 + h^2}$, the integration extends over the portion of the upper $\xi - \tau$ plane bounded between the two hyperbolas $C_0\tau = \sqrt{\xi^2 + h^2}$ and $t - \tau = \sqrt{(x - \xi)^2 + y^2}/C_0$. The two hyperbolas are tangent to each other when $s_2(x, y, t) = C_0 t - \sqrt{x^2 + (y + h)^2} = 0$. Hence S_2 vanishes on and outside the Mach cone $s_2 = 0$. When the point $(x > 0, y, t)$ lies in between the cone $s_2 = 0$ and

$$s_2'(x, y, t) = C_0 t - h - \sqrt{x^2 + y^2} = 0,$$

the area bounded between the two hyperbolas lies entirely in the first quadrant of the $\xi - \tau$ plane. Hence the disturbance due to the corner, 0, is confined by the Mach cone $s_2' = 0$, (Fig. 1b).

Similarly, we can express S_2', the additional disturbance pressure behind the barrier ($x < 0$), as a Possio integral,

$$S_2'(x, y, t) = -\frac{1}{\pi} \iint \frac{g_1(-|\xi|, \tau)\, d\xi\, d\tau}{\sqrt{(t - \tau)^2 - [(x - \xi)^2 + y^2]/C_0^2}} \tag{4}$$

S_2' vanishes on and outside the Mach cone $s_2 = 0$, while the disturbance due to the corner, 0, is confined by the cone $s_2' = 0$, (Fig. 1b).

Since S_1 is not an even function of x, S_2 and S_2' are in general different from each other. On the y-axis

$$S_{2x}(0^+, y, t) = S_{2x}'(0^-, y, t) = 0$$

$$S_2(0^+, y \geqq C_0 t - h, t) = S_2'(0^-, y \geqq C_0 t - h, t) = 0$$

but $$S_2(0^+, y < C_0 t - h, t) \neq S_2'(0^-, y < C_0 t - h, t).$$

Thus a difference in pressure in front of and behind the barrier is created.

During the interval $h/C_0 \leqq t \leqq 2h/C_0$, the total disturbance pressure is represented by $S_1 + S_2$ for $x > 0$ and $S_1 + S_2'$ for $x < 0$, and there is no discontinuity in pressure across the Mach cones $s_1 = 0$, $s_2 = 0$, and $s_2' = 0$.

(c) *The Third Problem.* When $t \geqq (2h/C_0)$, the difference in $S_2(0^+, y, t)$ and $S_2'(0^-, y, t)$ would create a discontinuity in pressure, $\Delta P_2 = S_2 - S_2'$, along the portion of the y-axis where $C_0 t - h > y > h$. The additional disturbance, which will be confined inside the Mach cone,

$$s_3(x, y, t) = C_0 t - 2h - \sqrt{(y - h)^2 + x^2} = 0,$$

should be such that the total disturbance pressure, p, and the x- and y-components of the disturbance velocity, i.e., u and v, are continuous across the portion of the $y - t$ plane where $y > h$ and $t \geqq 2h/C_0$. The conditions on velocity components, i.e.,

$$u(0^+, y, t) = u(0^-, y, t)$$

$$v(0^+, y, t) = v(0^-, y, t)$$

imply

$$u_t(0^+, y, t) = u_t(0^-, y, t)$$

and $$v_t(0^+, y, t) = v_t(0^-, y, t)$$

which, in turn, imply the continuity of the first derivatives of p on account of the linearized equations of motion.

Let S_3 and S_3' represent the additional disturbance pressure for $x > 0$ and $x < 0$ respectively. They will both obey the simple wave equation and the following boundary conditions:

$$S_{3x}(0^+, y < h, t) = S_{3x}'(0^-, y < h, t) = 0 \tag{5a}$$

$$S_{3x}(0^+, y > h, t) = S_{3x}'(0^-, y > h, t) \tag{5b}$$

$$S_3(0^+, y > h, t) - S_3'(0^-, y > h, t) = -\Delta P_2 \tag{5c}$$

and $$S_3 = S_3' = 0 \text{ on and outside the Mach cone } s_3 = 0 \tag{5d}$$

This problem can be solved by using the method developed by Evvard [7, 8] and Ward [9]. Due to the boundary conditions (5), S_3 and S_3' can be written

in the form of Possio integrals

$$S_3(x > 0, y, t) = -S_3'(x < 0, y, t)$$

$$= -\frac{1}{\pi}\iint \frac{g_3(\eta, \tau)\, d\eta\, d\tau}{\sqrt{[t - \tau - (2h/C_0)]^2 - \{x^2 + (y - h - \eta)^2\}/C_0^2}} \tag{6}$$

The integration extends over the domain in the first quadrant of the $\eta - \tau$ plane bounded by the hyperbola $C_0(t - \tau) - \sqrt{x^2 + (y - h - \eta)^2} - 2h = 0$, and the straight line $C_0\tau = \eta$, where

$$g_3(\eta > 0, \tau > 0) = S_{3z}(0, \eta + h, \tau + 2h/C_0) = S_{3z}'(0^-, \eta + h, \tau + 2h/C_0)$$

is, as yet, an unknown function.

To determine g_3, we apply the boundary condition (5c) which gives:

$$-\frac{1}{\pi}\iint \frac{g_3(\eta, \tau)\, d\eta\, d\tau}{\sqrt{[t - \tau - (2h/C_0)]^2 - (y - \eta - h)^2/C_0^2}} = -\tfrac{1}{2}\Delta p_2(y > h, t) \tag{7}$$

The integration is taken over the portion of the first quadrant of the $\eta - \tau$ plane bounded by the straight lines, $C_0\tau = \eta$ and

$$C_0(t - \tau - 2h/C_0) = \pm (y - \eta - h).$$

Following Evvard and Ward's method, we transform the integral by putting

$$\alpha = C_0 t - y - h, \beta = C_0 t + y - 3h$$

$$\alpha' = C_0\tau - \eta \quad \text{and} \quad \beta' = C_0\tau + \eta$$

Eq. (7) becomes

$$\frac{1}{\pi}\int_0^\alpha d\alpha' \left(\int_{\alpha'}^\beta \frac{G_3(\alpha', \beta')\, d\beta'}{\sqrt{(\alpha - \alpha')(\beta - \beta')}}\right) = f_3(\alpha, \beta) \tag{8}$$

where $$G_3(\alpha', \beta') = g_3\left(\frac{\beta' - \alpha'}{2}, \frac{\beta' + \alpha'}{2C_0}\right)$$

and $$f_3(\alpha, \beta) = \Delta p_2\left(\frac{\beta - \alpha + 2h}{2}, \frac{\beta + \alpha + 4h}{2C_0}\right) \qquad \text{with } \beta > \alpha.$$

This equation is an integral equation of Abel's type [8, 9] for the contents of the brackets, and the solution is

$$\int_\alpha^\beta \frac{G_3(\alpha, \beta')}{\sqrt{(\beta - \beta')}}\, d\beta' = \int_0^\alpha \frac{f_{3\alpha}(\alpha', \beta)\, d\alpha}{\sqrt{\alpha - \alpha'}} \tag{9}$$

Eq. (9) is another Abel's integral equation for G_3 and its solution is

$$G_3(\alpha, \beta) = \frac{1}{\pi}\frac{\partial}{\partial\beta}\int_\alpha^\beta \frac{d\beta'}{\sqrt{\beta - \beta'}}\int_0^\alpha \frac{f_{3\alpha}(\alpha', \beta')\, d\alpha'}{\sqrt{\alpha - \alpha'}} \tag{10}$$

Knowing $G_3(= g_3)$, we can evaluate S_3 and S_3' from eq. (6). During the interval $2h/C_0 \leqq t \leqq 3h/C_0$, the total disturbance pressures in front of and behind the structure are $S_1 + S_2 + S_3$ and $S_1 + S_2' + S_3'$ respectively.

When $t \geqq 3h/C_0$, the disturbances, S_3(for $x > 0$) and S_3'(for $x < 0$), confined by the Mach cone $s_3 = 0$ are reflected by the ground. The additional disturbances due to reflection, S_4(for $x > 0$) and S_4'(for $x < 0$) are related to S_{3y} and S_{3y}' the same as S_2 and S_2' to S_{1y} in the second problem. S_4 and S_4' vanish on and outside the Mach cone $s_4(x, y, t) = C_0t - 2h - \sqrt{x^2 + (y + h)^2} = 0$ while the influence of corner "0" (Fig. 1) is confined by the Mach cone $s_4'(x, y, t) = C_0t - 3h - \sqrt{x^2 + y^2} = 0$. When $t \geqq 4h/C_0$, the disturbances S_4 and S_4' interfere with each other over the barrier ($x = 0$, $y > h$). This interference causes an additional disturbance, S_5(for $x > 0$) and S_5'(for $x < 0$). They are related to S_4 and S_4' the same as S_3 and S_3' to S_2 and S_2' in the third problem. S_5 and S_5' will be confined by the Mach cone

$$S_5(x, y, t) = C_0t - 4h - \sqrt{x^2 + (y - h)^2} = 0.$$

It is clear that after the instant, $t = (2n - 1)h/C_0$, the additional disturbances created during the preceding interval $(2n - 2)h/C_0 \leqq t \leqq (2n - 1)h/C_0$, are reflected from the ground. Due to the reflection the additional disturbances S_{2n}(for $x > 0$) and S_{2n}'(for $x < 0$), which are solutions to the second problem, interfere with each other over the barrier after the instant $t = 2nh/C_0$ and cause the additional disturbances S_{2n+1} and S_{2n+1}', which are solutions to the third problem.

Based upon the preceding reasoning, we can compose the expression for the disturbance pressure for any instant, t, from the solutions to the preceding three problems as follows:

(a) During the interval $(2n - 1)h/C_0 \leqq t \leqq 2nh/C_0$

$$P = \begin{cases} S_1 + \sum_{i=2}^{2n} S_i & \text{for } x > 0 \\ S_1 + \sum_{k=2}^{2n} S_k' & \text{for } x < 0 \end{cases}$$

S_{2n}(or S_{2n}') is a solution to the second problem and is related to S_{2n-1}(or S_{2n-1}') the same as S_2(or S_2') to S_1.

(b) During the interval $(2nh)/C_0 \leqq t \leqq (2n + 1)h/C_0$

$$P = \begin{cases} S_1 + \sum_{k=2}^{2n+1} S_k & \text{for } x > 0 \\ S_1 + \sum_{k=2}^{2n+1} S_k' & \text{for } x < 0 \end{cases}$$

S_{2n+1}(or S_{2n+1}') is a solution to the third problem and is related to S_{2n}(or S_{2n}') the same as S_3(or S_3') to S_2(or S_2').

Across all the Mach cones,

$$S_{2n}(x, y, t) = C_0t - (2n - 2)h - \sqrt{x^2 + (y + h)^2} = 0,$$

$$S_{2n}'(x, y, t) = C_0t - (2n - 1)h - \sqrt{x^2 + y^2} = 0,$$

and $$S_{2n+1}(x, y, t) = C_0t - 2nh - \sqrt{x^2 + (y - h)^2} = 0,$$

the disturbance pressure in the first approximation is continuous. However, second order (ϵ^2) shocks may exist in the neighborhood of the Mach cones [10, 11, 12].

3. Rectangular Barriers of Finite Width ($L/h = 0(1)$). For the convenience in explanation, we assume $2h > L > h$, but actually the solution can also cover the case of $L \leqq h$ and $L \geqq 2h$. As in the problem of thin barrier, the solution can be composed from the solutions to the following three problems:

(a) *The First Problem.* Fig. (3a) represents the shock pattern during the interval $0 \leqq t \leqq h/C_0$. The disturbance pressure will be the same if we consider the barrier as an infinite wedge of 90° wedge angle. The solution S_1 can be obtained in a similar manner as that for the thin barrier with slight modification due to the difference in boundary conditions.

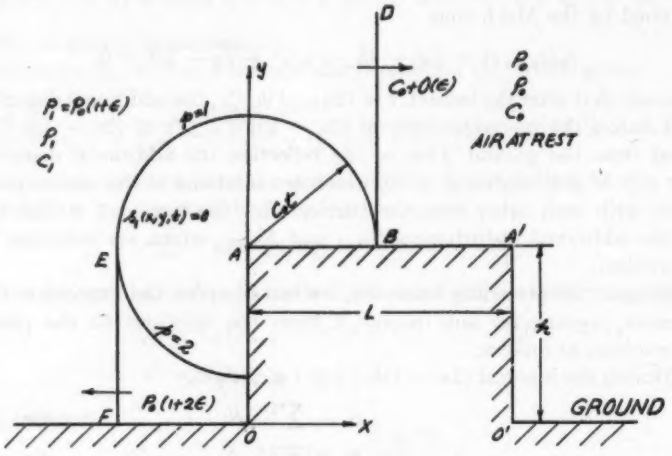


FIG 3a DURING THE INTERVAL $0<t<h/C_0$

(b) *The Second Problem.* When $t > h/C_0$, an additional disturbance pressure S_2' is created behind the barrier. S_2' is related to S_1 by eq. (4) (Fig. 3b).

(c) *The Third Problem.* When $t \geqq L/C_0$, the Mach cone

$$S_1' = C_0 t - \sqrt{x^2 + (y-h)^2} = 0$$

is passing over the corner A' (Fig. 3c). The solution S_1 is in general different from zero along the line $y = h$, although its derivative S_{1y} vanishes there. A discontinuity in pressure, $\Delta p = S_1$, would be created along the line $A'B$ (Fig. 3c). The additional disturbance which will be confined inside the Mach cone,

$$s_3(x, y, t) = C_0 t - L - \sqrt{(x-L)^2 + (y-h)^2} = 0$$

should be such that the discontinuity of pressure along the line $A'B$ is removed.

Let S_3 and S_3' represent the additional disturbance pressure in regions I and II respectively. They will both obey the simple wave equation, eq. (1), and the boundary conditions as shown in Fig. (3c).

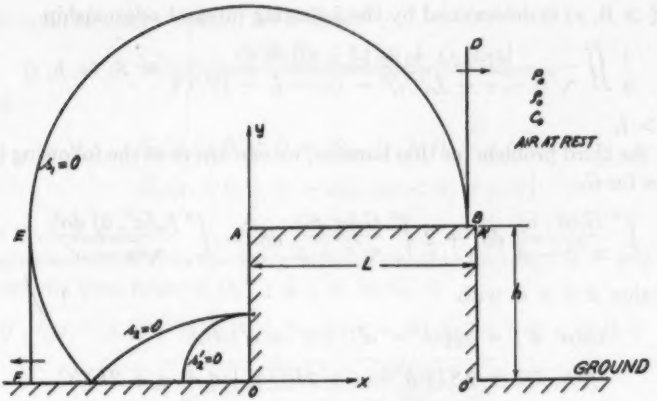


FIG 3b. DURING THE INTERVAL
$h/C_0 < t < L/C_0$

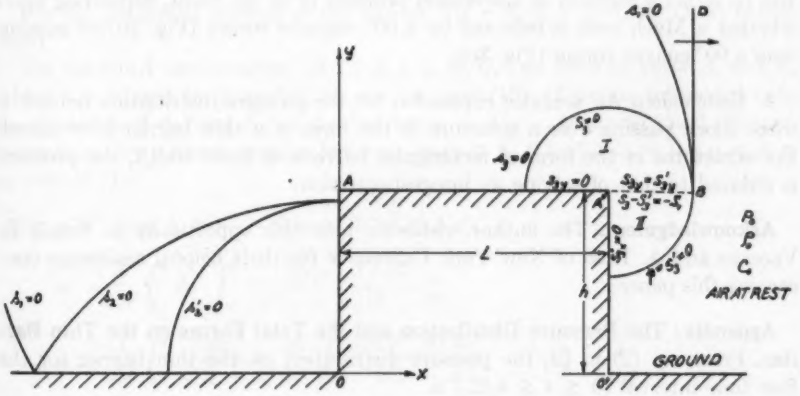


FIG. 3c During The Interval $L/C_0 < t < 2h/C_0$

The Possio integral solutions for S_3 and S_3' are

$$S_3 = -\frac{1}{\pi}\iint \frac{g_3(\xi,\tau)\,d\xi\,d\tau}{\sqrt{(t-\tau-L/C_0)^2-[(x-L-\xi)^2+y^2]/C_0^2}}$$

and

$$S_3' = +\frac{1}{\pi}\iint \frac{g_3(|\xi|,\tau)\,d\xi\,d\tau}{\sqrt{(t-\tau-L/C_0)^2-[(x-L-\xi)^2+y^2]/C_0^2}}$$

where $$g_3(\xi<0,\tau)=0$$

and $g_3(\xi > 0, \tau)$ is determined by the following integral relationship

$$\frac{1}{\pi}\iint \frac{[g_3(\xi, \tau) + g_3(|\xi|, \tau)]\, d\xi\, d\tau}{\sqrt{(t - \tau - L/C_0)^2 - (x - L - \xi)^2/C_0^2}} = S_1(x, h, t)$$

with $x > L$.

As in the third problem for thin barriers, we can arrive at the following integral equation for G_3,

$$\int_0^\alpha \frac{G_3(\beta', \alpha)}{\sqrt{\beta - \beta'}}\, d\beta' + 2\int_\alpha^\beta \frac{G_3(\alpha, \beta')}{\sqrt{\beta - \beta'}}\, d\beta' = \int_0^\alpha \frac{f_{3a}(\alpha', \beta)\, d\alpha'}{\sqrt{\alpha - \alpha'}} \cdots \qquad (11)$$

for all value of $\beta > \alpha$, with

$$G_3(\alpha', \beta') = g_3[\tfrac{1}{2}(\beta' - \alpha'), \tfrac{1}{2}(\beta' + \alpha')/C_0]$$

and

$$f_3(\alpha, \beta) = 2S_1[\tfrac{1}{2}(\beta - \alpha + 2L), h, \tfrac{1}{2}(\beta + \alpha + 2L)/C_0$$

The author has not been able to obtain a solution of eq. (11) in a closed form.

As in the thin barrier problem, after the instant of $t = (mL + nh)/C_0$, where n and m are two positive integers, an addition disturbance will be produced. It will be either a solution to the second problem or to the third, depending upon whether a Mach cone is reflected by a 90° concave corner (Fig. 3b) or passing over a 90° convex corner (Fig. 3c).

4. Conclusion. An analytic expression for the pressure distribution behind a weak shock passing over a structure in the form of a thin barrier is obtained. For structures in the form of rectangular barriers of finite width, the problem is reduced to that of solving an integral equation.

Acknowledgment. The author wishes to take this opportunity to thank R. Vaccaro and A. Ross of New York University for their helpful assistance concerning this paper.

Appendix: The Pressure Distribution and the Total Forces on the Thin Barrier. From eq. (2) in §2, the pressure distribution on the thin barrier for the first time interval $(0 \leqq t \leqq h/C_0)$ is

$$p(x = 0^-, 0 \leqq y \leqq h, t) = 2 - p(x = 0^+, 0 \leqq y \leqq h, t)$$

$$= 1 + \frac{1}{\pi}\arctan\frac{2\sqrt{(h - y)(C_0 t - h + y)}}{C_0 t - 2(h - y)} \qquad \text{(A-1)}$$

Since the solution $-2 + S_1(x < 0, -y, t)$, which is a solution to the simple wave equation, vanishes on the portion of the Mach cone $s_2(x, y, t) = 0$ inside the second quadrant of $x - y$ plane, and also fulfills the boundary conditions for S_2' on $x = 0$, and $y = 0$, we can conclude that

$$\text{and} \quad \begin{cases} S_2'(x < 0, y, t) = S_1(x < 0, -y, t) - 2 & \text{for } s_2 \geqq 0 \\ S_2' = 0 & \text{for } s_2 \leqq 0 \end{cases} \qquad \text{(A-2)}$$

without evaluating the Possio integral in eq. (3).

Based upon the same reason, we obtain

$$S_2(x > 0, y, t) = S_1(x > 0, -y, t) \tag{A-3}$$

and also

$$\left.\begin{aligned} S_{2n}(x > 0, y, t) &= S_{2n-1}(x > 0, -y, t) \\ S'_{2n}(x < 0, y, t) &= S'_{2n-1}(x < 0, -y, t) \end{aligned}\right\} \tag{A-4}$$

where $n = 2, 3, 4 \cdots$.

The pressure distribution on the portion of the thin barrier, $C_0 t - h \geqq y \geqq 0$, for the second time interval $(h/C_0 \leqq t \leqq 2h/C_0)$ is

$$\begin{aligned} p(x = 0^-, C_0 t - h \geqq y \geqq 0, t) &= 2 - p(x = 0^+, C_0 t - h \geqq y \geqq 0, t) \\ &= \frac{1}{\pi} \arctan \frac{2\sqrt{(h - y)(C_0 t - h + y)}}{C_0 t - 2(h - y)} \\ &\quad + \frac{1}{\pi} \arctan \frac{2\sqrt{(h + y)(C_0 t - h - y)}}{C_0 t - 2(h + y)} \end{aligned} \tag{A-4}$$

For the remainder portion of the barrier, $C_0 t - h \leqq y \leqq h$, eq. (A-1) is still valid.

For the third time interval, $2h/C_0 \leqq t \leqq 3h/C_0$, we have to know S_3 and S'_3 which are related to S_2 and S'_2 by eqs. (6) and (10). If we are interested only on the value of S_3 (and S'_3) on the barrier ($x = 0$, $y < h$) eqs. (6) and (10) can be replaced by a direct relation between $\Delta p_2(y > h, t)$ and $S_3(x = 0^+, y < h, t)$ as follows [9]:

$$\begin{aligned} S_3(x = 0^+, y \leqq h, t) &= -S'_3(x = 0^-, y \leqq h, t) \\ &= -\frac{1}{2\pi} \int_0^\beta d\alpha' \int_{\alpha'}^\beta \frac{G_3(\alpha', \beta')\, d\beta}{\sqrt{(\alpha - \alpha')(\beta - \beta')}} \\ &= -\frac{1}{2\pi} \int_0^\beta \frac{d\alpha'}{\sqrt{\alpha - \alpha'}} \frac{\partial}{\partial \alpha'} \int_0^{\alpha'} \frac{f_3(\alpha'', \beta)\, d\alpha''}{\sqrt{\alpha - \alpha''}} \\ &= -\frac{1}{2\pi} \frac{1}{\sqrt{\alpha - \beta}} \int_0^\beta \frac{f_3(\alpha'', \beta)}{\sqrt{\beta - \alpha''}}\, d\alpha'' \\ &\quad + \frac{1}{4\pi} \int_0^\beta \frac{d\alpha}{(\alpha - \alpha')^{3/2}} \int_0^{\alpha'} \frac{f_3(\alpha'', \beta)\, d\alpha''}{\sqrt{\alpha' - \alpha''}} \\ &= -\frac{\sqrt{\alpha - \beta}}{2\pi} \int_0^\beta \frac{f_3(\alpha'', \beta)\, d\alpha''}{(\alpha - \alpha'')\sqrt{\beta - \alpha''}} \\ &= -\frac{\sqrt{\alpha - \beta}}{2\pi} f_3(\beta, \beta) \arctan \sqrt{\frac{\beta}{\alpha - \beta}} \\ &\quad + \frac{\alpha - \beta}{2\pi} \int_0^\beta \frac{f_3(\beta, \beta) - f_3(\alpha'', \beta)}{(\alpha - \alpha'')\sqrt{\beta - \alpha''}}\, d\alpha'' \end{aligned} \tag{A-5}$$

Similarly, the values of $S_{2n+1}(= -S'_{2n+1})$ on the barrier $(x = 0, y < h)$ can be related directly to

$$\Delta p_{2n}(y > h, t)[= -S'_{2n}(x = 0^-, y > h, t) + S_{2n}(x = 0^+, y > h, t) = -2S'_{2n}]$$

as $S_3(= -S'_3)$ to Δp_2 in eq. (A-5).

Thus, we have succeeded to show that the only difficulty involved in obtaining the pressure distribution on the barrier for any instant is to evaluate repeatedly the definite integrals of the type in eq. (A-5). This can always be done by numerical method (Simpson's one-third rule).

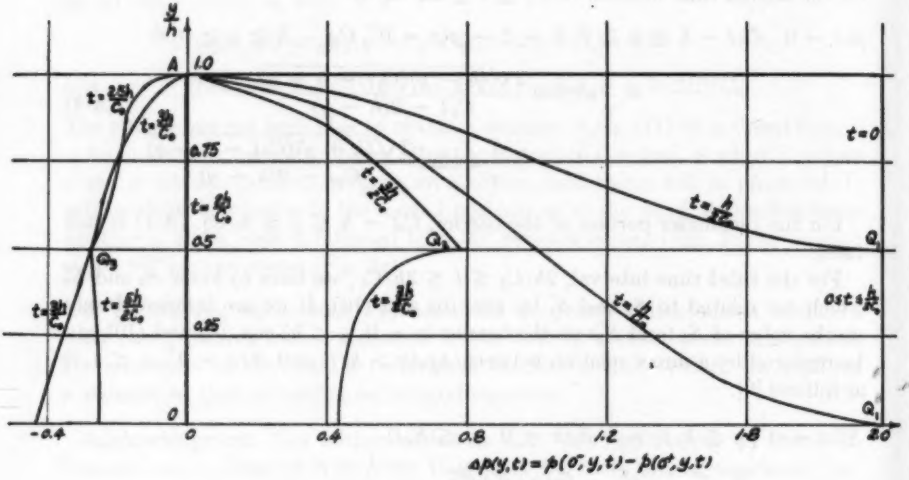


FIG. 4 THE DIFFERENCE IN PRESSURE IN FRONT OF AND BEHIND THE THIN BARRIER

The pressure difference behind and in front of the thin barrier,

$$p(0^-, 0 \leqq y \leqq h, t) - p(0, 0 \leqq y \leqq h, t) = \Delta p(0 \leqq y \leqq h, t)$$

for the instances, $t = h/(2C_0)$, $t = h/C_0$, $t = 3h/2C_0$, $t = 2h/C_0$, $t = 5h/(2C_0)$ and $t = 3h/C_0$, is shown in Fig. 4. The curves are continuous, but their slopes have discontinuities at points Q_1, Q_2, and Q_3. This indicates that the first derivatives of the disturbance pressure $p(x, y, t)$ are discontinuous across the Mach cones $s_1 = 0$, $s_2 = 0$, and $s_3 = 0$. The disturbance pressure behind and in front of the thin barrier can be obtained from their difference by the following simple relations.

$$p(0^-, 0 \leqq y \leqq h, t) = 1 + \tfrac{1}{2}\Delta p(0 \leqq y \leqq h, t)$$

and $$p(0^+, 0 \leqq y \leqq h, t) = 1 - \tfrac{1}{2}\Delta p(0 \leqq y \leqq h, t)$$

Fig. 5 shows the variation of the total force acting on the thin barrier for the time interval, $0 \leqq t \leqq 3h/C_0$.

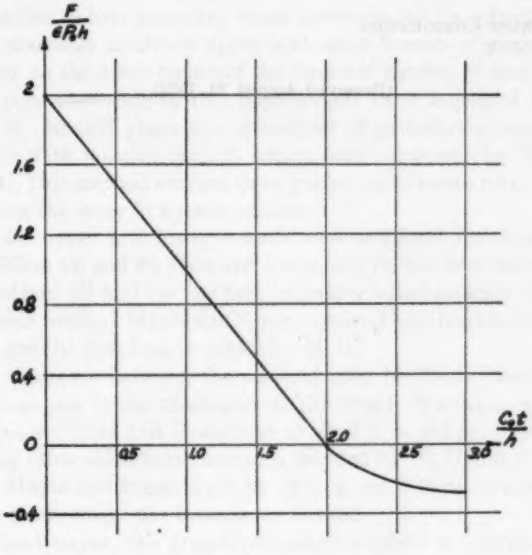


FIG. 5 THE TOTAL FORCE ACTING ON THE THIN BARRIER

REFERENCES

1. Bargmann, V., On Nearly Glancing Reflection of Shocks. AMP Rep. 108, 2R, 1945.
2. Lighthill, M. J., The Diffraction of Blast—I. Proc. Roy. Soc., Series A, **198**, 1949.
3. Lighthill, M. J., The Diffraction of Blast—II. Proc. Roy. Soc., Series A, **200**, 1950.
4. Ting, L., Aerodynamics of Blasts. Doctorate Dissertation, New York University, April 1951. Part of it was published under the same title by Lu Ting and H. F. Ludloff in the Journal of the Aero. Sc. **19**, no. 5, 317–328, May 1952.
5. Keller, J. B. and Blank, A., Diffraction and Reflection of Pulses by Wedges and Corners. Communication on Pure and Applied Mathematics **4**, no. 1, June 1951.
6. Courant, R. and Friedrichs, K. O., Supersonic Flow and Shockwaves. New York: Interscience Publishers, Inc., 1948, p. 154.
7. Evvard, J. C., Distribution of Wave Drag and Lift in the Vicinity of Wing Tips at Supersonic Speeds. N. A. C. A., T. N. no. 1382, July 1947.
8. Evvard, J. C. and Turner, R. L., Theoretical Lift Distribution and Upwash Velocity for Thin Wings at Supersonic Speeds. N. A. C. A., T. N. no. 1484, Nov. 1947.
9. Ward, G. N., Supersonic Flow Past Thin Wings. Quart. Journ. Mech. and Appl. Math. **11**, pt. 2, 1949.

10. Lighthill, M. J., A Technique for Rendering Approximate Solutions to Physical Problems Uniformly Valid. Phil. Mag., Series 7, **40**, 1179–1201, Nov. 1949.
11. Lighthill, M. J., The Shock Strength in Supersonic Conical Fields. Phil. Mag., Series 7, **40**, 1202–1221, Nov. 1949.
12. Ting, L., The Shock Strength in a Two Dimensional Nonsteady Flow. Jour. Aero. Sc. **19**, no. 5, 351–352, May, 1952.

Foster Wheeler Corporation
New York, N. Y.

(Received August 22, 1952)

BOUNDS FOR SOLUTIONS OF BOUNDARY VALUE PROBLEMS IN ELASTICITY

By K. Washizu

1. Introduction. When boundary value problems can be solved only approximately, it is desirable to obtain upper and lower bounds of the solution. Since Trefftz's paper on the lower bound of the torsional rigidity of bars was published [1], many papers answering to this requirement have appeared in the field of elasticity [2, 3]. Several years ago, a method of geometrical representation in function space with positive definite metric was presented by W. Prager and J. L. Synge [4]. This method enables us to grasp approximate methods intuitively and to estimate the error in a mean square.

Formulae for upper and lower bounds were obtained by Greenberg for the Dirichlet problem [5] and by Diaz and Greenberg for the first biharmonic boundary value problem [6] and for the first boundary value problem of elasticity [7]. Using a different method Maple and Synge obtained also bounds for the Dirichlet problem and for the problems in elasticity [8–11].

The chief difference between the method used by Diaz-Greenberg and that by Maple-Synge lies in the treatment of the Green functions, which have not integrable squares. Diaz and Greenberg succeed in avoiding divergent integrals by introducing other subsidiary functions denoted by W, U and V in the present paper, while Maple and Synge do it by cutting out a finite circular or spherical region at the point where the bounds are sought.

In the present paper, the Diaz-Greenberg's method is followed, and unified formulae for bounds of solutions in problems of elasticity, including two- and three-dimensional problems, bending of plates and torsion of bars, are represented geometrically in function space. The Green functions above mentioned are called elementary solutions in this paper, being denoted by φ. They have definite physical meanings and are easily obtainable in many cases. Several examples of elasticity are taken, but the same method will be applicable to other similar problems.

2. Definition of the three dimensional stress problem. Let the problem be defined by the following equations:

(1) Equations of equilibrium

$$\partial\sigma_x/\partial x + \partial\tau_{yx}/\partial y + \partial\tau_{zx}/\partial z = 0, \text{ etc.} \tag{1}$$

where body forces are assumed to be absent for the sake of simplicity.

(2) Conditions of compatibility, i.e. conditions that strain components are derived from u, v and w such that

$$\varepsilon_x = \partial u/\partial x, \text{ etc.} \qquad \gamma_{xy} = \partial u/\partial y + \partial v/\partial x, \text{ etc.} \tag{2}$$

(3) Stress-strain relations

$$E\varepsilon_x = \sigma_x - \nu(\sigma_y + \sigma_z), \text{ etc.} \qquad G\gamma_{xy} = \tau_{xy}, \text{ etc.} \tag{3}$$

(4) Boundary conditions

$$\begin{pmatrix} X_r \\ Y_r \\ Z_r \end{pmatrix} \equiv \begin{pmatrix} \sigma_x & \tau_{yx} & \tau_{zx} \\ \tau_{xy} & \sigma_y & \tau_{zy} \\ \tau_{xz} & \tau_{yz} & \sigma_z \end{pmatrix} \begin{pmatrix} l \\ m \\ n \end{pmatrix} = \begin{pmatrix} \bar{X}_r \\ \bar{Y}_r \\ \bar{Z}_r \end{pmatrix} \qquad \text{on } S_1, \quad (4a)$$

$$u = \bar{u}, \qquad v = \bar{v}, \qquad w = \bar{w} \qquad \text{on } S_2 \quad (4b)$$

where l, m and n are the direction cosines of the normal drawn outwards on the boundary, and S_1 and S_2 denote the parts of the boundary where the force and displacement are specified, respectively. For the sake of brevity, we will treat simply-connected bodies only.

3. Set of functions. According to the method propounded by H. J. Greenberg and W. Prager, let us consider a set of $(\sigma_x, \sigma_y, \cdots, \tau_{xy})$ as a vector in the function space and adopt vectorial notations defined in the following:

S is the exact solution of the original problem;
S' satisfies the conditions with respect to stresses only, namely, (1) and (4a);
I'_p satisfies (1) and homogeneous boundary conditions on S_1, namely, $X_r = 0$, $Y_r = 0$ and $Z_r = 0$, $(p = 1, 2, \cdots, m)$;
S'' satisfies the conditions with respect to displacements only, namely (2) and (4b);
I''_q satisfies (2) and homogeneous boundary conditions on S_2, namely, $u = 0$, $v = 0$ and $w = 0$, $(q = 1, 2, \cdots, n)$;
φ is an elementary solution with an isolated singularity at the point (ξ, η, ζ), it satisfies the field conditions only, namely, (1) and (2);
W satisfies (1) and (2), and $\varphi + W$ satisfies homogeneous boundary conditions of stresses on S_1 and of displacements on S_2, respectively, (it is the regular part of a certain Green function);
U satisfies (1) and the same boundary conditions as W on S_1;
V satisfies (2) and the same boundary conditions as W on S_2;

It is assumed that the relations (3) hold in every case.

4. Definition of a scalar product. Let two arbitrary vectors be denoted by N and N', and their components of stress and strain be represented by $(\sigma_x, \sigma_y, \cdots, \tau_{xy})$, $(\varepsilon_x, \varepsilon_y, \cdots, \gamma_{xy})$ and $(\sigma'_x, \sigma'_y, \cdots, \tau'_{xy})$, $(\varepsilon'_x, \varepsilon'_y, \cdots, \gamma'_{xy})$, respectively. The scalar product of the two vectors in the function-space is defined by

$$(N, N') \equiv \iiint_V (\sigma_x \varepsilon'_x + \sigma_y \varepsilon'_y + \sigma_z \varepsilon'_z + \cdots + \tau_{xy} \gamma'_{xy})\, dv, \qquad (5)$$

the integrals being taken throughout the body. From (5) the following relations

are obtained immediately:

$$(N, N') = (N', N), \qquad (N, N) \equiv \| N \|^2 > 0$$

$$| (N, N') | \leq \| N \| \cdot \| N' \| \qquad \text{Schwarz's inequality}$$

If we replace N' in (5) by S'', and denote the displacement components of S'' by (u'', v'', w''), then we get

$$(S'', N) = \iint_{S_1+S_2} [u'' X_\nu + v'' Y_\nu + w'' Z_\nu] \, dS$$

$$- \iiint_V \left[u'' \left(\frac{\partial \sigma_x}{\partial x} + \frac{\partial \tau_{yx}}{\partial y} + \frac{\partial \tau_{zx}}{\partial z} \right) + v'' \left(\frac{\partial \tau_{xy}}{\partial x} + \frac{\partial \sigma_y}{\partial y} + \frac{\partial \tau_{zy}}{\partial z} \right) \right. \tag{6}$$

$$\left. + w'' \left(\frac{\partial \tau_{xz}}{\partial x} + \frac{\partial \tau_{yz}}{\partial y} + \frac{\partial \sigma_z}{\partial z} \right) \right] dv$$

$$\equiv [S'', N]_1 + [S'', N]_2 - \iiint_V [u''(\cdots) + v''(\cdots) + w''(\cdots)] \, dv,$$

where the brackets $[\ ]_1$ and $[\ ]_2$ denote the surface integrals on S_1 and S_2, respectively. As in the above, the bracket is so defined that it is built with the displacement components of the first vector and the stress components of second one. The operation denoted by the bracket is generally non-commutative. Writing $\bar{S}$ for the given surface value, we find that

$$(S'', S') = [S'', \bar{S}]_1 + [\bar{S}, S']_2, \qquad (S'', I'_p) = [\bar{S}, I'_p]_2$$

$$(I''_q, S') = [I''_q, \bar{S}]_1, \qquad (I'_p, I''_q) = 0,$$

$$\| I'_p \pm I''_q \|^2 = \| I'_p \|^2 + \| I''_q \|^2$$

The last two relations mean that I'_p and I''_q are orthogonal in the function space. The geometrical representations of these relations are illustrated in fig. 1.

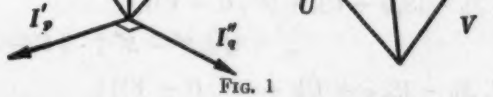


FIG. 1

5. Fundamental inequalities. Let us consider the scalar product of φ and S in two ways, namely,

$$(\varphi, S) = [S, \varphi]_1 + [S, \varphi]_2 + A(\xi, \eta, \zeta)$$

$$= [S, \varphi]_1 + [\bar{S}, \varphi]_2 + A \tag{7a}$$

and

$$(\varphi, S) = [\varphi, S]_1 + [\varphi, S]_2 + A'(\xi, \eta, \zeta)$$
$$= [\varphi, \bar{S}]_1 + [\varphi, S]_2 + A', \tag{7b}$$

where A and A' are certain functions depending on the singularity of φ. Since we have

$$[S, \varphi]_1 = -[S, U]_1 = -(S, U) + [S, U]_2 = -(S, U) + [\bar{S}, U]_2 \tag{8a}$$

and

$$[\varphi, S]_2 = -[V, S]_2 = -(V, S) + [V, S]_1 = -(V, S) + [V, \bar{S}]_1, \tag{8b}$$

we get from (7a) and (7b),

$$A - A' = [\varphi + V, \bar{S}]_1 - [\bar{S}, \varphi + U]_2 + (S, U - V)$$
$$= [\varphi + V, S]_1 - [\bar{S}, \varphi + U]_2 + (S', U - V) + (S - S', U - V) \tag{9a}$$
$$= [\varphi + V, \bar{S}]_1 - [\bar{S}, \varphi + U]_2 + (S'', U - V) + (S - S'', U - V) \tag{9b}$$
$$= [\varphi + V, \bar{S}]_1 - [\bar{S}, \varphi + U]_2 + (\tfrac{1}{2}(S' + S''), U - V) +$$
$$(S - \tfrac{1}{2}(S' + S''), U - V). \tag{9c}$$

By the Schwarz's inequality already mentioned we get

$$|(S - S', U - V)| \leqq \|S - S'\| \cdot \|U - V\| \leqq \|S' - S''\| \cdot \|U - V\|$$
$$|(S - S'', U - V)| \leqq \|S - S''\| \cdot \|U - V\| \leqq \|S' - S''\| \cdot \|U - V\|$$

and

$$|(S - \tfrac{1}{2}(S' + S''), U - V)| \leqq \tfrac{1}{2}\|S' - S''\| \cdot \|U - V\|.$$

Consequently, if we put

$$A - A' \equiv P,$$

then upper and lower bounds of th value of P are obtainable in terms of S', S'', φ, U, and V, namely,

$$|P - \{[\varphi + V, \bar{S}]_1 - [\bar{S}, \varphi + U]_2 + (S', U - V)\}|$$
$$\leqq \|S' - S''\| \cdot \|U - V\| \tag{10a}$$
$$|P - \{[\varphi + V, \bar{S}]_1 - [S, \varphi + U]_2 + (S'', U - V)\}|$$
$$\leqq \|S' - S''\| \cdot \|U - V\| \tag{10b}$$

and

$$|P - \{[\varphi + V, \bar{S}]_1 - [\bar{S}, \varphi + U]_2 + \tfrac{1}{2}(S' + S'', U - V)\}|$$
$$\leqq \tfrac{1}{2}\|S' - S''\| \cdot \|U - V\|. \tag{10c}$$

The accuracy of the above bounds can be improved by putting

$$S' = S'_0 + \sum_{p=1}^{m} a_p I'_p, \qquad U = U_0 + \sum_{p=1}^{m} c_p I'_p,$$
$$S'' = S''_0 + \sum_{q=1}^{n} b_q I''_q, \qquad V = V_0 + \sum_{q=1}^{n} d_q I''_q,$$

where

$$S'_0 \in S', \quad S''_0 \in S'', \quad U_0 \in U, \quad \text{and} \quad V_0 \in V,$$

and determining the arbitrary constants a_p, b_q, c_p and d_q so that $\| S' - S'' \|$ and $\| U - V \|$ are reduced as small as possible.

6. Some examples of elementary solutions. Some examples are listed in the table 1. The singular point (ξ, η, ζ) being taken at the origin of the coordinates, P is the value at the origin.

TABLE 1

	Displacement components defined by φ	P
I.	$u_\varphi = (3 - 4\nu)\frac{1}{r} + \frac{x^2}{r^3}$	
	$v_\varphi = \frac{xy}{r^3}$	$16(1-\nu)\pi G \cdot u$
	$w_\varphi = \frac{xz}{r^3}$	
II.	$u_\varphi = (1 - 4\nu)\frac{x}{r^3} + 3\frac{x^3}{r^5}$	
	$v_\varphi = -\frac{y}{r^3} + 3\frac{x^2 y}{r^5}$	$16(1-\nu)\pi G \cdot \frac{\partial u}{\partial x}$
	$w_\varphi = -\frac{z}{r^3} + 3\frac{x^2 z}{r^5}$	
III.	$u_\varphi = (1 - 2\nu)\frac{x}{r^3} + 3\frac{x^3}{r^5}$	
	$v_\varphi = -(1 - 2\nu)\frac{y}{r^3} + 3\frac{x^2 y}{r^5}$	$8(1-\nu)\pi \cdot \sigma_x$
	$w_\varphi = -(1 - 2\nu)\frac{z}{r^3} + 3\frac{x^2 z}{r^5}$	
IV.	$u_\varphi = (1 - 2\nu)\frac{y}{r^3} + 3\frac{x^2 y}{r^5}$	
	$v_\varphi = (1 - 2\nu)\frac{x}{r^3} + 3\frac{xy^2}{r^5}$	$8(1-\nu)\pi \cdot \tau_{xy}$
	$w_\varphi = 3\frac{xyz}{r^5}$	

The expressions I correspond to the solution of an infinite body subjected to a concentrated load in the direction of the x-axis [12]. The expressions II are derived from I [12]. The former ones are same as those given by Diaz and Greenberg [7] though based on the scalar product (5) defined differently. The expressions III and IV correspond to the solutions of an infinite body with a concentrated dislocation. The expressions III are derived from the solution of an infinite body, which has a circle of discontinuity with respect to the displacement u in the (y, z) plane, by reducing the radius of the circle to zero, keeping the product (the dislocation strength) $\times$ (the area of the circle) constant [13]. The expressions IV are the solution of an infinite body with a point of discontinuity with respect to the displacement v in the (y, z) plane, obtained by the same limiting process as above.

The free Green vectors for stress and displacement were obtained by Synge—the so-called Maple-Synge method [10]—by using the property that the dilatation or the sum of the principal stresses is a harmonic function and applying the mean theorem of the potential theory. The free Green vectors obtained by the said author are similar in their forms to the elementary solutions listed above. The latters defined by the field conditions (1) and (2) are, however, more restricted in their definition and more definite in their physical meanings than the formers.

So far, some elementary solutions of three dimensional problems have been obtained. As the elementary solution has to be selected according to the problem, let us take up in the following some special problems, *viz.*, two dimensional stress problems, bending of thin plates, and torsion of bars.

7. Two dimensional stress problem. Let the problem be defined by the following fundamental equations:

(1) Equations of equilibrium

$$\partial\sigma_x/\partial x + \partial\tau_{yx}/\partial y = 0, \qquad \partial\tau_{xy}/\partial x + \partial\sigma_y/\partial y = 0.$$

(2) Conditions of compatibility, i.e. conditions that strain components are derived from u and v such that

$$\varepsilon_x = \partial u/\partial x, \qquad \varepsilon_y = \partial v/\partial y, \qquad \gamma_{xy} = \partial u/\partial y + \partial v/\partial x.$$

(3) Stress-strain relations

$$E\varepsilon_x = \sigma_x - \nu\sigma_y, \qquad E\varepsilon_y = -\nu\sigma_x + \sigma_y, \qquad G\gamma_{xy} = \tau_{xy}.$$

(4) Boundary conditions

$$\begin{pmatrix} X_\nu \\ Y_\nu \end{pmatrix} \equiv \begin{pmatrix} \sigma_x & \tau_{yx} \\ \tau_{xy} & \sigma_y \end{pmatrix} \begin{pmatrix} l \\ m \end{pmatrix} = \begin{pmatrix} \bar{X}_\nu \\ \bar{Y}_\nu \end{pmatrix} \qquad \text{on } S_1,$$

$$u = \bar{u}, \qquad v = \bar{v} \qquad \text{on } S_2.$$

Some examples of φ and P are listed in the table 2, in which $x = r\cos\theta$ and $y = r\sin\theta$.

TABLE 2

	Airy's stress function defined by φ	P
I.	$\frac{1-\nu}{2} x \log r - y\theta$	$2\pi u$
II.	$-\log r + \frac{x^2}{r^2}$	$\frac{4\pi}{E}\sigma_x$
III.	$\frac{xy}{r^2}$	$\frac{4\pi}{E}\tau_{xy}$

As it can be easily seen, the expression I corresponds to the solution of an infinite plane subjected to a concentrated force acting at the origin in the direction of the x-axis. The expressions II and III are derived from the dislocation functions $x \log r$ and $y \log r$, respectively. When necessary, the table can be appended with the functions for orthogonally aeolotropic plates [14].

8. Bending of thin plates. Let the problem be defined by the following equations:

(1) Equations of equilibrium

$$\partial^2 M_{11}/\partial x^2 + 2\partial^2 M_{12}/\partial x \partial y + \partial^2 M_{22}/\partial y^2 + p = 0. \tag{11}$$

(2) Conditions of compatibility, i.e. conditions that strain components are derived from w such that

$$\kappa_{11} = -\partial^2 w/\partial x^2, \qquad \kappa_{22} = -\partial^2 w/\partial y^2, \qquad \kappa_{12} = -\partial^2 w/\partial x \partial y. \tag{12}$$

(3) Stress-strain relations

$$M_{11} = D(\kappa_{11} + \nu\kappa_{22}), \qquad M_{22} = D(\nu\kappa_{11} + \kappa_{22}), \qquad M_{12} = D(1-\nu)\kappa_{12}. \tag{13}$$

(4) Boundary conditions

$$\text{mechanical B. C.} \quad G = f(s),\ N + \partial H/\partial s = g(s) \quad \text{on } C_1, \tag{14a}$$

$$\text{geometrical B. C.} \quad w = h(s),\ \partial w/\partial n = k(s) \quad \text{on } C_2, \tag{14b}$$

where $\partial/\partial n$ means the derivative with respect to the normal n drawn outwards on the boundary, and G and $N + \partial H/\partial s$ mean the bending moment and shearing force on the boundary, respectively. The values of w and p are taken positive when they are in the z direction.

It is defined that

S is the exact solution;

S' satisfies (11) and (14a);

I'_p satisfies the homogeneous equation of equilibrium

$$\partial^2 M_{11}/\partial x^2 + 2\partial^2 M_{12}/\partial x \partial y + \partial^2 M_{22}/\partial y^2 = 0; \tag{15}$$

and homogeneous boundary conditions on C_1 ;

S'' satisfies (12) and (14b);
I_q'' satisfies (12) and homogeneous boundary conditions on C_2 ;
φ is an elementary solution and satisfies (12) and (15);
W satisfies (12) and (15), and $\varphi + W$ satisfies homogeneous boundary conditions on C_1 and C_2 ;
U satisfies (15) and the same boundary conditions as W on C_1 ;
V satisfies (12) and the same boundary conditions as W on C_2 .

Let two arbitrary vectors be denoted by N and N', and their stress and strain components be denoted by (M_{11}, M_{22}, M_{12}), $(\kappa_{11}, \kappa_{22}, \kappa_{12})$ and $(M'_{11}, M'_{22}, M'_{12})$, $(\kappa'_{11}, \kappa'_{22}, \kappa'_{12})$, respectively. The scalar product is defined by

$$(N, N') \equiv \iint (M_{11}\kappa'_{11} + M_{22}\kappa'_{22} + 2M_{12}\kappa'_{12})\, dx\, dy.$$

If we replace N' in the above expression by S'' and denote the displacement of S'' by w'', then we obtain

$$(S'', N) = -\oint\left(G\frac{\partial w''}{\partial n} + H\frac{\partial w''}{\partial t}\right) ds + \oint w''N\, ds - \iint w''\left(\frac{\partial^2 M_{11}}{\partial x^2} + 2\frac{\partial^2 M_{12}}{\partial x \partial y} + \frac{\partial^2 M_{22}}{\partial y^2}\right) dx\, dy, \tag{16}$$

where

$$G \equiv M_{11}l^2 + 2M_{12}lm + M_{22}m^2,$$

$$H \equiv (M_{22} - M_{11})lm + M_{12}(l^2 - m^2),$$

$$N \equiv (\partial M_{11}/\partial x + \partial M_{12}/\partial y)l + (\partial M_{12}/\partial x + \partial M_{22}/\partial y)m,$$

and l and m are the direction cosines of the normal n; moreover $\partial/\partial t$ means the derivative with respect to the tangent t, which is so drawn that n, t and z constitute a right hand system. When t is taken so as to coincide with ds, (16) is transformed into

$$(S'', N) = \oint\left[w''\left(N + \frac{\partial H}{\partial s}\right) - \frac{\partial w''}{\partial n}G\right] ds - \iint w''\left(\frac{\partial^2 M_{11}}{\partial x^2} + 2\frac{\partial^2 M_{12}}{\partial x \partial y} + \frac{\partial^2 M_{22}}{\partial y^2}\right) dx\, dy. \tag{17}$$

The relation (17) is used in the subsequent formulations, which are almost the same as in the preceding sections except some modifications needed for the inequalities, namely,

$$(\varphi, S) = [S, \varphi]_1 + [\bar{S}, \varphi]_2 + A(\xi, \eta), \tag{18a}$$

$$(\varphi, S) = [\varphi, \bar{S}]_1 + [\varphi, S]_2 + A'(\xi, \eta) + \iint p\varphi \, dx \, dy, \tag{18b}$$

$$[S, \varphi]_1 = -[S, U]_1 = -(S, U) + [S, U]_2 = -(S, U) + [\bar{S}, U]_2, \tag{19a}$$

$$[\varphi, S]_2 = -[V, S]_2 = -(V, S) + [V, \bar{S}]_1 + \iint pV \, dx \, dy, \tag{19b}$$

where φ and V in the integrals mean the displacements.
Hence, the following formulae for the bounds of $P(\equiv A - A')$ are obtained:

$$\begin{aligned} P &= [\varphi + V, \bar{S}]_1 - [\bar{S}, \varphi + U]_2 + \iint p(\varphi + V) \, dx \, dy + (S, U - V) \\ &= [\varphi + V, \bar{S}]_1 - [\bar{S}, \varphi + U]_2 + \iint p(\varphi + V) \, dx \, dy \\ &\qquad + (S', U - V) + (S - S', U - V) \end{aligned} \tag{20a}$$

$$\begin{aligned} &= [\varphi + V, \bar{S}]_1 - [\bar{S}, \varphi + U]_2 + \iint p(\varphi + V) \, dx \, dy \\ &\qquad + (S'', U - V) + (S - S'', U - V), \end{aligned} \tag{20b}$$

and

$$\left| P - \left\{ [\varphi + V, \bar{S}]_1 - [\bar{S}, \varphi + U]_2 + \iint p(\varphi + V) \, dx \, dy + (S', U - V) \right\} \right| \leqq \| S' - S'' \| \cdot \| U - V \|, \tag{21a}$$

$$\left| P - \left\{ [\varphi + V, \bar{S}]_1 - [\bar{S}, \varphi + U]_2 + \iint p(\varphi + V) \, dx \, dy + (S'', U - V) \right\} \right| \leqq \| S' - S'' \| \cdot \| U - V \|. \tag{21b}$$

Some examples of elementary solutions are listed in the table 3.

TABLE 3

Displacement defined by φ	P	Displacement defined by φ	P
I. $r^2 \log r$	$8\pi D \cdot w$	III. $-(1+\nu)\log r - (1-\nu)\frac{x^2}{r^2}$	$4\pi M_{11}$
II. $-(2\log r + 1)x$	$8\pi D \cdot \frac{\partial w}{\partial x}$	IV. $-(1-\nu)\frac{xy}{r^2}$	$4\pi M_{12}$

These functions are taken from the theory of bending of thin plates [15]. Bounds for the expression I are equivalent to those obtained by Diaz and Greenberg [6]. When necessary, the table can be appended with the functions for orthogonally aeolotropic plates [16].

9. Torsion of bars. Lastly the torsion problem of bars is dealt with; the problem is defined as:

(1) Equations of equilibrium

$$\partial\tau_{xz}/\partial x + \partial\tau_{yz}/\partial y = 0.$$

(2) Condition of compatibility, i.e. condition that strain components are derived from w such that

$$\gamma_{xz} = \partial w/\partial x - y\Theta, \qquad \gamma_{yz} = \partial w/\partial y + x\Theta.$$

(3) Stress-strain relations

$$G\gamma_{xz} = \tau_{xz}, \qquad G\gamma_{yz} = \tau_{yz}.$$

(4) Boundary condition

$$\tau_{xz}\, l + \tau_{yz}\, m = 0,$$

where l and m are the direction cosines of the normal drawn outwardly at the boundary, and Θ is the twist angle per unit length. Alternatively the problem may be written in terms of the stress-function:

(1)′ Equation of equilibrium

$$\tau_{xz} = \partial\phi/\partial y, \qquad \tau_{yz} = -\partial\phi/\partial x. \tag{22}$$

(2)′ Condition of compatibility

$$\partial^2\phi/\partial x^2 + \partial^2\phi/\partial y^2 + 2G\,\Theta = 0. \tag{23}$$

(4)′ Boundary condition

$$\phi = 0. \tag{24}$$

It is defined that

S is the exact solution;
S' satisfies (23);
I'_p is a plane harmonic function;
S'' satisfies (24);
I''_q satisfies (24);
φ is an elementary solution being a plane harmonic function;
W is a plane harmonic function, and $\varphi + W = 0$ on the boundary;
U is a plane harmonic function;
V has the same boundary value as W on the boundary;

Take two vectors N and N', of which the torsion stress functions are denoted by ψ and ψ', respectively. The scalar product is defined by

$$(N, N') \equiv \iint\left(\frac{\partial\psi}{\partial x}\frac{\partial\psi'}{\partial x} + \frac{\partial\psi}{\partial y}\frac{\partial\psi'}{\partial y}\right)dx\,dy,$$

which is equivalent to (5) and can be transformed into

$$(N, N') = \oint \psi\frac{\partial\psi'}{\partial n}\,ds - \iint \psi\left(\frac{\partial^2\psi'}{\partial x^2} + \frac{\partial^2\psi'}{\partial y^2}\right)dx\,dy, \tag{26}$$

n denoting the direction of the normal drawn outwardly. Some inequalities can be written by putting

$$(\varphi, S) = A(\xi, \eta), \tag{27}$$

$$(\varphi, S) = [\varphi, S] + \iint 2G\Theta\cdot\varphi\, dx\, dy + A'(\xi, \eta), \tag{28}$$

$$[\varphi, S] = -[V, S] = -(V, S) + \iint 2G\Theta\cdot V\, dx\, dy, \tag{29}$$

where φ and V in the integrals are the torsion-stress-functions. Hence, formulae for bounds of $P(=A - A')$ are obtained:

$$P = -(V, S) + \iint 2G\Theta(\varphi + V)\, dx\, dy$$

$$= (S', U - V) + \iint 2G\Theta(\varphi + V)\, dx\, dy + (S - S', U - V) \tag{30a}$$

$$= (S'', U - V) + \iint 2G\Theta(\varphi + V)\, dx\, dy + (S - S'', U - V) \tag{30b}$$

therefore

$$\left| P - \left\{(S', U - V) + \iint 2G\Theta(\varphi + V)\, dx\, dy\right\}\right| \leqq \| S' - S'' \| \cdot \| U - V \|, \tag{31a}$$

and

$$\left| P - \left\{(S'', U - V) + \iint 2G\Theta(\varphi + V)\, dx\, dy\right\}\right| \leqq \| S' - S'' \| \cdot \| U - V \|. \tag{31b}$$

Some examples of elementary solutions are listed in the table 4.

TABLE 4

Torsion-stress-function defined by φ		P
I.	$-\log r$	$2\pi\phi$
II.	x/r^2	$2\pi\partial\phi/\partial x$
III.	y/r^2	$2\pi\partial\phi/\partial y$

The expression I corresponds to the solution w/Θ = real part of $(-i \log z)$, where $z = x + iy$ and $i = \sqrt{-1}$. The solution has a constant dislocation with respect to the displacement w around the origin. The expressions II and III are derived from I. Bounds for these expressions are equivalent to those obtained previously by Greenberg [5].

10. Final remarks. It was tried in the present note to build a common road leading to the bounds in search along the line already drawn by the former authors whose works were referred to in particular with regard to the examples of the singular solutions written in the tables. It may be concluded that upper and lower bounds of a certain scalar, vector, or tensor quantity in the problem of elasticity are theoretically obtainable by using these singular solutions coupled with auxiliary functions. There are difficulties, however, in finding these auxiliary functions, especially U and V, so that the required conditions are satisfied; moreover, modified methods are needed when we seek bounds at points on the boundary. The treatment of these difficulties will be a problem in the future.

The author wishes to express his sincere thanks to Prof. A. Ono for his kind inspection of this paper. He wishes also to express his gratitude to the referee for some of the references listed following.

REFERENCES

(1) E. Trefftz, Proc. 2nd. Int. Congr. Appl. Mech., Zürich 1926, p. 131.
(2) K. Friedrichs, Göttinger Nachrichten, 1929, p. 13.
(3) N. M. Basu, Phil. Mag., **10**, 886, 1930.
(4a) W. Prager and J. L. Synge, Quart. Appl. Math., **5**, 241, 1947.
(4b) H. J. Greenberg and W. Prager, Am. J. Math., **70**, p. 749, 1948.
(5) H. J. Greenberg, J. Math. Phys., **27**, 161, 1948.
(6) J. B. Diaz and H. J. Greenberg, J. Math. Phys., **27**, 193, 1948.
(7) J. B. Diaz and H. J. Greenberg, Quart. Appl. Math., **6**, 326, 1948.
(8) J. L. Synge, Proc. Roy. Soc., **191**, 447, 1949.
(9) C. G. Maple, Quart. Appl. Math., **8**, 213, 1950.
(10) J. L. Synge, Proc. Roy. Irish. Acad., **53**, 41, 1950.
(11) J. L. Synge, Proc. Roy. Soc., **208**, 170, 1951.
(12) A. E. H. Love, A Treatise on the Mathematical Theory of Elasticity. Cambridge University Press, 1934, p. 185.
(13) J. M. Burgers, Proc. Kon. Ned. Acad. Wet., **42**, 293, 1939.
(14) S. Moriguti, Trans. Symposium Appl. Mech. First Faculty of Engineering, Univ. of Tokyo, **2**, 54, 1947 (in Japanese).
(15) K. Washizu, Trans. Japan Soc. Mech. Eng., **18**, 41, 1952 (in Japanese).
(16) K. Washizu, Proc. 1st Japan Nat. Congr. Appl. Mech., 1951, p. 157.

Faculty of Engineering
University of Tokyo

(Received January 2, 1952)

ON A VARIATIONAL THEOREM FOR FINITE ELASTIC DEFORMATIONS[1]

By Eric Reissner

1. Introduction. In the following we formulate a variational theorem of the theory of finite elastic deformations, which is characterised by the fact that the Euler equations of the variational problem consist of the differential equations of equilibrium *and* the stress-displacement relations and for which stress *and* displacement boundary conditions are natural boundary conditions.

A corresponding result for infinitesimal deformations has been indicated earlier [1, 2, 3]. It is found here that generalization to finite deformations is quite direct if one works with the notion of stress introduced by Trefftz [4] and further developed by Kappus [5].[2]

On the basis of the theorem for three-dimensional elasticity it is possible to state analogous relations for beams, plates and shells. We do this here for the problem of finite bending of plates. The appropriate variational equation is such that from it one may obtain by specialization results previously given by Kirchhoff [6], Marguerre [7] and Wang [8].

2. Differential equations for finite elastic deformations. Let x_m be the cartesian coordinates of a point of a body before deformation and let $\mathbf{r}$ be the corresponding radius vector, written in the form,

$$\mathbf{r} = \sum x_m \, \mathbf{i}_m \tag{1}$$

Let $\mathbf{r} + \mathbf{u}$ be the radius vector to the same material point after deformation and write the displacement vector $\mathbf{u}$ in the form

$$\mathbf{u} = \sum u_m \, \mathbf{i}_m \tag{2}$$

An infinitesimal rectangular parallelepiped with edge vectors $(\partial \mathbf{r}/\partial x_j)\, dx_j = \mathbf{i}_j \, dx_j$ in the undeformed state is deformed into a parallelepiped which in general is not rectangular. Its edge vectors are $[\partial(\mathbf{r} + \mathbf{u})/\partial x_j]\, dx_j = [\mathbf{i}_j + (\partial \mathbf{u}/\partial x_j)]\, dx_j = \mathbf{g}_j \, dx_j$ where

$$\mathbf{g}_j = \sum \left(\delta_{jm} + \frac{\partial u_m}{\partial x_j} \right) \mathbf{i}_m \tag{3}$$

Components of finite strain e_{jk} may be defined in terms of the lattice vectors $\mathbf{g}_j$ by means of the relations $2e_{jk} = \mathbf{g}_j \cdot \mathbf{g}_k - \delta_{jk}$, or

$$e_{jk} = \tfrac{1}{2} \sum \delta_{mk} \frac{\partial u_m}{\partial x_j} + \delta_{mj} \frac{\partial u_m}{\partial x_k} + \frac{\partial u_m}{\partial x_j} \frac{\partial u_m}{\partial x_k} \tag{4}$$

[1] The present paper is a report on work done under the sponsorship of the Office of Naval Research under Contract N5-ori-07834 with Massachusetts Institute of Technology.

[2] This was first reported in a colloquium at Brown University on March 14, 1952. The same result was found independently by B. Frays de Veubeke.

Let $\mathfrak{s}_j$ be the force, per unit of undeformed area, acting on an element of area which before deformation was perpendicular to the x_j-direction. Components of (pseudo) stress are defined, following Trefftz [4], by writing

$$\mathfrak{s}_j = \sum s_{jk}\,\mathfrak{g}_k \tag{5}$$

Introduction of (5) and (3) into the force equilibrium equation $\sum \partial \mathfrak{s}_j/\partial x_j + \mathfrak{f} = 0$ leads to the following three scalar equilibrium equations

$$\sum\sum \frac{\partial}{\partial x_j}\left[s_{jk}\left(\delta_{km} + \frac{\partial u_m}{\partial x_k}\right)\right] + f_m = 0 \tag{6}$$

The conditions of moment equilibrium for the deformed parallelepipedon are of the form

$$s_{jk} = s_{kj} \tag{7}$$

We shall assume in the following that the components of body force intensity f_m may be written in terms of a function $\Psi(x_1, \ldots ; u_1, \ldots)$ in the form

$$f_m = \frac{\partial \Psi}{\partial u_m} \tag{8}$$

The system of equations (6), (7) and (8) is completed by the stress strain relations which for the present purposes are taken in the form

$$e_{jk} = \frac{\partial W}{\partial s_{jk}} \tag{9}$$

where W is a given function of its arguments s_{mn}.

For the formulation of boundary conditions we require a component representation for surface forces. Let p_m be the x_m-component of surface force, per unit of undeformed surface area, acting on an element of area which before deformation was perpendicular to a vector $\mathfrak{v}$. The condition of force equilibrium for an infinitesimal tetrahedron before deformation with sides perpendicular to the vectors $\mathfrak{v}$ and $\mathfrak{i}_m$ gives

$$p_m = \sum\sum \cos(\nu, x_j) s_{jk}\left(\delta_{km} + \frac{\partial u_m}{\partial x_k}\right) \tag{10}$$

Let S_1 be that part of the surface of the body where stresses are prescribed and S_2 that part of the surface where displacements are prescribed. The boundary value problems of the theory of elasticity with finite deformation which we consider here consist of the differential equations (6) and (9) together with the boundary conditions

$$\begin{aligned} &\text{on } S_1 : p_m = \bar{p}_m \\ &\text{on } S_2 : u_m = \bar{u}_m \end{aligned} \tag{11}$$

where $\bar{p}_m$ and $\bar{u}_m$ are given functions.

3. The variational theorem. The following theorem will be proved. *The state of stress and displacement which satisfies the differential equations of equilibrium and the stress displacement relations in the interior of the body, and the conditions of prescribed stress on the part S_1 and of prescribed displacement on the part S_2 of the surface of the body, is determined by the variational equation*

$$\delta\left\{\iiint [\sum\sum s_{jk}\, e_{jk} - W - \Psi]\, dV - \iint_{S_1} [\sum \bar{p}_m u_m]\, dS - \iint_{S_2} [\sum (u_m - \bar{u}_m) p_m]\, dS\right\} = 0. \tag{12}$$

In order to see the validity of (12) it is noted that according to the rules of the calculus of variations and in view of the definitions of the functions W and Ψ the variational equation (12) is equivalent to,

$$\iiint \left[\sum\sum \left(e_{jk} - \frac{\partial W}{\partial s_{jk}}\right)\delta s_{jk} + s_{jk}\,\delta e_{jk} - \sum \frac{\partial \Psi}{\partial u_m}\,\delta u_m\right] dV - \iint_{S_1} [\sum \bar{p}_m\,\delta u_m]\, dS - \iint_{S_2} [\sum (u_m - \bar{u}_m)\delta p_m + p_m\,\delta u_m]\, dS = 0 \tag{13}$$

From (4) we have for the variations of the components of strain e_{jk},

$$\delta e_{jk} = \tfrac{1}{2}\sum \delta_{mk}\frac{\partial \delta u_m}{\partial x_j} + \delta_{mj}\frac{\partial \delta u_m}{\partial x_j} + \frac{\partial u_m}{\partial x_j}\frac{\partial \delta u_m}{\partial x_k} + \frac{\partial u_m}{\partial x_k}\frac{\partial \delta u_m}{\partial x_j} \tag{14}$$

With (14) we obtain by suitable integration by parts,

$$\begin{aligned}\iiint \sum\sum s_{jk}\,\delta e_{jk}\, dV = {} & -\tfrac{1}{2}\iiint \sum\sum\sum \left\{\frac{\partial}{\partial x_j}\left[\left(\delta_{mk} + \frac{\partial u_m}{\partial x_k}\right)s_{jk}\right]\right. \\ & \left. + \frac{\partial}{\partial x_k}\left[\left(\delta_{mj} + \frac{\partial u_m}{\partial x_j}\right)s_{jk}\right]\right\}\delta u_m\, dV \\ & + \tfrac{1}{2}\iint \sum\sum\sum s_{jk}\left\{\left(\delta_{mk} + \frac{\partial u_m}{\partial x_k}\right)\cos(\nu, x_j)\right. \\ & \left. + \left(\delta_{mj} + \frac{\partial u_m}{\partial x_j}\right)\cos(\nu, x_k)\right\}\delta u_m\, dS\end{aligned} \tag{15}$$

Since $s_{jk} = s_{kj}$ equation (15) may also be written in the form

$$\begin{aligned}\iiint \sum\sum\sum s_{jk}\,\delta e_{jk}\, dV = {} & -\iiint \sum\sum\sum \frac{\partial}{\partial x_j}\left[\left(\delta_{mk} + \frac{\partial u_m}{\partial x_k}\right)s_{jk}\right]\delta u_m\, dV \\ & + \iint \sum\sum\sum \cos(\nu, x_j)\, s_{jk}\left(\delta_{mk} + \frac{\partial u_m}{\partial x_k}\right)\delta u_m\, dS\end{aligned} \tag{16}$$

The surface integral in (16) may be rewritten in terms of p_m according to (10).

Introduction of (16) and (13) gives then

$$\iiint \Big\{ \sum\sum \Big(e_{jk} - \frac{\partial W}{\partial s_{jk}} \Big) \delta s_{jk} - \Big\{ \sum\sum\sum \frac{\partial}{\partial x_j} \Big[\Big(\delta_{mk} + \frac{\partial u_m}{\partial x_k} \Big) s_{jk} \Big] + f_m \Big\} \delta u_m \Big\} dV + \iint_{S_1} \sum (p_m - \bar{p}_m) \delta u_m \, dS - \iint_{S_2} \sum (u_m - \bar{u}_m) \delta p_m \, dS = 0 \tag{17}$$

Since in the interior of the body we have δs_{jk} and δu_m arbitrary, on the surface S_1 δu_m arbitrary and on S_2 δp_m arbitrary there follows that the variational equation (17) is equivalent to the differential equations (6) and (9) and to the boundary conditions (11). This proves the theorem.

4. Stress strain relations. For many problems for which non-linearity due to finite deformations must be considered although the strains themselves are very small, it is adequate to assume linear relations between the components of pseudo stress s_{jk} and the components of finite strain e_{jk}. If in addition it is assumed that the material is isotropic then we may write

$$E e_{jk} = (1 + \nu) s_{jk} - \nu \delta_{jk} \sum s_{mm} \tag{18}$$

where E is the modulus of elasticity and ν is Poisson's ratio. For a material with the stress strain relations (18) the function W introduced through equation (9) is of the form

$$W = (1/2E)\{(1 + \nu)\sum\sum s_{jk}^2 - \nu(\sum\sum \delta_{jk} s_{jk})^2\} \tag{19}$$

As an example of stress strain relations for finite strain we may consider the stress strain relations for an incompressible isotropic material which has been termed neo-Hookean. According to Rivlin [9] the relations between components of principal strain and principal stress for this material are of the form

$$t_N = \tfrac{1}{3} E(1 + 2e_N) + p, \qquad N = I, II, III \tag{20}$$

where p is of the nature of a hydrostatic pressure the magnitude of which follows by application of the condition of incompressibility,

$$(1 + 2e_I)(1 + 2e_{II})(1 + 2e_{III}) = 1 \tag{21}$$

The determination of p from (20) and (21) requires the solution of a cubic equation.

In view of the definition of the components of pseudo stress s_{mn} the relation between the component of principal stress t_I and the component of principal pseudo stress s_I is of the form

$$t_I \sqrt{(1 + 2e_{II})(1 + 2e_{III})} = s_I \sqrt{1 + 2e_I}, \tag{22}$$

with corresponding relations between t_N and s_N. For an incompressible material

equation (22) may be written alternately as

$$t_I = s_I(1 + 2e_I) \tag{23}$$

Therewith the stress strain relations (20) assume the form

$$s_N = \tfrac{1}{3}E + p/(1 + 2e_N) \tag{24}$$

Combination of (24) with the incompressibility relation (21) next gives the following explicit expression for p

$$p = -\sqrt[3]{(\tfrac{1}{3}E - s_I)(\tfrac{1}{3}E - s_{II})(\tfrac{1}{3}E - s_{III})} \tag{25}$$

and with this the stress strain relations (24) may be written as

$$2e_N = \frac{\sqrt[3]{(\tfrac{1}{3}E - s_I)(\tfrac{1}{3}E - s_{II})(\tfrac{1}{3}E - s_{III})}}{\tfrac{1}{3}E - s_N} - 1 \tag{26}$$

Since $e_N = \partial W/\partial s_N$ we have then for the function W

$$W = -\tfrac{3}{2}\sqrt[3]{(\tfrac{1}{3}E - s_I)(\tfrac{1}{3}E - s_{II})(\tfrac{1}{3}E - s_{III})} - \tfrac{1}{2}(s_I + s_{II} + s_{III}) + \tfrac{1}{2}E \tag{27}$$

In order to obtain relations between e_{jk} and s_{jk} from (27) we introduce in (27) the invariants

$$\begin{aligned} I_1 &= s_{11} + s_{22} + s_{33} = s_I + s_{II} + s_{III} \\ I_2 &= s_{11}s_{22} + s_{11}s_{33} + s_{22}s_{33} - s_{12}s_{21} - s_{13}s_{31} - s_{23}s_{32} \\ &= s_I s_{II} + s_I s_{II} + s_{II} s_{III} \\ I_3 &= \begin{vmatrix} s_{11} & s_{12} & s_{13} \\ s_{21} & s_{22} & s_{23} \\ s_{31} & s_{32} & s_{33} \end{vmatrix} = s_I s_{II} s_{III} \end{aligned} \tag{28}$$

In terms of these,

$$W = -\tfrac{1}{2}E\sqrt[3]{1 - \frac{3}{E}I_1 + \frac{9}{E^2}I_2 - \frac{27}{E^3}I_3} - \tfrac{1}{2}I_1 + \tfrac{1}{2}E \tag{29}$$

If (29) is developed according to powers of s/E the result is, up to third-degree terms,

$$W = \frac{I_1^2 - 3I_2}{2E} - \frac{5I_1^3 + 18I_1I_2 - 27I_3}{6E^2} + \cdots \tag{30}$$

Retention of only second degree terms leads back to the linear relations (18) between s_{jk} and e_{jk}, with ν having the value $\frac{1}{2}$.

5. Finite bending of plates. The variational equation corresponding to (12) for finite bending of plates as discussed by Kirchhoff [6] and von Karman [10] is

of the following form,

$$\delta\Big\{\iint\Big[N_x\Big(\frac{\partial u}{\partial x}+\tfrac{1}{2}\Big(\frac{\partial w}{\partial x}\Big)^2\Big)+N_y\Big(\frac{\partial v}{\partial y}+\tfrac{1}{2}\Big(\frac{\partial w}{\partial y}\Big)^2\Big)+N_{xy}\Big(\frac{\partial u}{\partial y}+\frac{\partial v}{\partial x}+\frac{\partial w}{\partial x}\frac{\partial w}{\partial y}\Big)$$
$$-M_x\frac{\partial^2 w}{\partial x^2}-M_y\frac{\partial^2 w}{\partial y^2}-2M_{xy}\frac{\partial^2 w}{\partial x\partial y}$$
$$-\frac{1}{2Eh}(N_x^2+N_y^2-2\nu N_xN_y+2(1+\nu)N_{xy}^2)$$
$$-\frac{12}{2Eh^3}(M_x^2+M_y^2-2\nu M_xM_y+2(1+\nu)M_{xy}^2)-\Psi(x,y,u,v,w)\Big]\,dx\,dy \qquad (31)$$
$$-\int_{s_1}\Big[\bar N_{\nu\nu}u+\bar N_{\nu\nu}v+\bar Q_\nu w-\bar M_{\nu\nu}\frac{\partial w}{\partial x}-\bar M_{\nu\nu}\frac{\partial w}{\partial y}\Big]\,ds$$
$$-\int_{s_2}\Big[(u-\bar u)N_{\nu\nu}+(v-\bar v)N_{\nu\nu}+(w-\bar w)Q_\nu$$
$$-\Big(\frac{\partial w}{\partial x}-\frac{\partial\bar w}{\partial x}\Big)M_{\nu\nu}-\Big(\frac{\partial w}{\partial y}-\frac{\partial\bar w}{\partial y}\Big)M_{\nu\nu}\Big]\,ds\Big\}=0$$

In (31) u, v, w are components of displacement of points of the middle surface, N_x, N_y, N_{xy}, Q are stress resultants and M_x, M_y, M_{xy} stress couples, ν is the direction in the plane of the undeflected middle surface normal to its boundary, and

$$N_{\nu\nu}=\cos(\nu,x)N_x+\cos(\nu,y)N_{xy}$$
$$N_{\nu\nu}=\cos(\nu,x)N_{xy}+\cos(\nu,y)N_y \qquad (32)$$

and $M_{\nu\nu}$ and $M_{\nu\nu}$ are defined correspondingly. The expression for Q_ν is

$$Q_\nu=\cos(\nu,x)\Big[V_x+N_x\frac{\partial w}{\partial x}+N_{xy}\frac{\partial w}{\partial y}\Big]$$
$$+\cos(\nu,y)\Big[V_y+N_{xy}\frac{\partial w}{\partial x}+N_y\frac{\partial w}{\partial y}\Big] \qquad (33)$$

and Q_ν is a force in the direction of z, while V_x and V_y are transverse shear stress resultants in the direction of the normal to the deformed middle surface, given by $V_x=\partial M_x/\partial x+\partial M_{xy}/\partial y$ and $V_y=\partial M_{xy}/\partial x+\partial M_y/\partial y$. Note that $\delta\bar Q_\nu=0$ although $\bar Q_\nu$ is obtained from Q_ν by putting bars on V and N but not on w.

Elimination of all stress resultants and couples from (31) by means of the stress strain relations and comparison of only such displacements as satisfy all prescribed displacements boundary conditions, together with the assumption that $\Psi=Xu+Yv+pw$ leads to the minimum-potential-energy equation for finite bending of plates as given by Kirchhoff [6].

Introduction in (31) of an Airy stress function F for N_x, N_y, N_{xy}, together with the assumption that $\Psi=pw$, elimination of the displacement components u and v and comparison of only such states as satisfy the boundary condition for

N_x, N_y, and N_{xy} and w leads to a variational equation given by Marguerre [7]. The Euler equations of this variational equation are the two simultaneous differential equations for F and w given by von Karman [10].

Finally, by restricting admissible states to those satisfying *all* equilibrium equations, equation (31) may be reduced to a recent result by Wang [8].

Equation (31) may be generalized in such a way that the effect of transverse shear stress deformation is taken into account. To this end we must replace in (31) the expression $-M_x\partial^2 w/\partial x^2 - M_y\partial^2 w/\partial y^2 - 2M_{xy}\partial^2 w/\partial x\partial y$ by $M_x\partial\alpha/\partial x + M_y\partial\beta/\partial y + M_{xy}(\partial\alpha/\partial y + \partial\beta/\partial x)$ and make corresponding changes in the boundary integrals. Furthermore we must add in the double integral the terms $V_x(\alpha + \partial w/\partial x) + V_y(\beta + \partial w/\partial y) - [12(1 + \nu)/5h](V_x^2 + V_y^2)$. The quantities α and β are projections on the x, z and y, z-planes, respectively, of the angle which the deformed normal encloses with the z-axis.

REFERENCES

1. Reissner, E., *On a variational theorem in elasticity*. J. Math. and Physics **29**, 90–95 (1950).
2. Reissner, E., *Stress-strain relations in the theory of thin elastic shells*. J. Math. and Physics **31**, 109–119 (1952).
3. Reissner, E., *On Non-uniform torsion of cylindrical rods*. J. Math. and Physics **31**, 214–221 (1952).
4. Trefftz, E., *Uber die Ableitung der Stabilitätskriterien des elastischen Gleichgewichts aus der Elastizitätstheorie endlicher Deformationen*. Proc. III Intern. Congr. Appl. Mech., vol. III, 44–50 (1931).
5. Kappus, R., *Zur Elastizitätstheorie endlicher Verschiebungen*. Z. ang. Math. and Mech. **19**, 271–285, 344–361 (1939).
6. Kirchhoff, G., *Vorlesungen über Mechanik, Dreissigste Vorlesung*. Leipzig, Teubner 1876, 1883 and 1897.
7. Marguerre, K., *Die mittragende Breite der gedrückten Platte*. Luftfahrtforschung **14**, 121–128 (1937).
8. Wang, C. T., *Principle and application of complementary energy method for thin homogeneous and sandwich plates and shells with finite deformations*. N. A. C. A., T. N. 2620 (1952).
9. Rivlin, R. S., *Large elastic deformations of isotropic materials*. Phil. Trans. Roy. Soc. **240**, 459–490 (1948).
10. von Karman, T., *Festigkeitsprobleme im Maschinenbau*. Encyclopädie der mathematischen Wissenschaften **4**, 1910, p. 349.

Massachusetts Institute of Technology

(Received October 15, 1952)

A MATHEMATICAL BASIS FOR AN ERROR ANALYSIS OF DIFFERENTIAL ANALYZERS*

By K. S. Miller and F. J. Murray

1.1. Purpose and Literature. The purpose of this paper is to present a mathematical basis for a general error analysis for the solution of systems of ordinary differential equations by machine methods. Specifically, we shall concern ourselves with the effect of errors on the machine solutions. We show that the study of these effects for general non-linear systems can be referred to the solution of linear systems of ordinary differential equations, but we do this without "linearizing" or simplifying the given system.

Since we permit perturbations in the solution as given by the machine, our discussion is applicable to both continuous computers and digital machines using step by step methods. However, stability discussions for such a digital process are most conveniently given in terms of difference equations rather than differential equations and are not given here.

There is an extensive literature on digital solutions dealing with truncation and rounding errors: Brock and Murray [1], Chadaja [7], Duncan [10, 11, 12], Forsythe [13], Fricke [14], Gill [16], Huskey [18], Kirby [19], Loud [26], Miller [29], Murray [30], Papoulis [32], Rademacher [33], Todd [36], Turton [38], but an intensive analysis of this work is not appropriate here. One of our major objectives is to avoid the "linearization" which appears in these and in this sense our results can be regarded as supplementing this work.

The effect of errors on solutions obtained by means of continuous computers has been studied in the case in which the given problem involves a system of linear equations with constant coefficients, notably by Raymond [34] and Macnee [27, 28].

Our discussion is based to a certain extent on well known theories for the dependence of systems of ordinary differential equations on parameters. However, it was necessary to extend this theory in order to properly consider those errors which affect the order of the system. Order variations in systems of equations have been considered from other points of view by Coddington and Levinson [8], Friedrichs and Wasow [15], Gradstein [17], and Levinson [25].

We have also listed in our bibliography certain papers in Russian of which we have considered only reviews: Bruevič [2], Byhovskiĭ [4, 5, 6], Tihonov [37], Vasil'eva [40, 41]. These papers may contain material relevant to our present discussion.

1.2. Objectives. A mathematical theory of errors should provide a framework within which errors can be studied and their effects evaluated. The study of errors of individual components should be oriented to such a framework so that one can estimate their effects on the solution. Error studies for components are

* The preparation of this paper was assisted by funds of the Office of Naval Research.

highly desirable but their value depends upon a knowledge of how the solution will be affected by them.

The effect of component or other individual errors on the solution depends on the problem considered. One can readily verify that certain problems are much more sensitive to errors than others. Consequently, it is very desirable to understand and specify this sensitivity. This can only be done by methods having the generality of the present paper. Few conclusions can be drawn from specific examples, even when a reference solution of undoubted correctness is available.

Another and very important reason for such an error study is that it gives an insight into the mathematical or theoretical structure of the system of differential equations considered. One should appreciate the great technical advance represented by the differential analyzer. Without machine computations one is practically limited to linear systems with constant coefficients. However, it is also true that the structure of the latter is well known. To obtain the equivalent information for the general problems handled on differential analyzers, one needs additional theoretical structure, part of which is indicated by this error analysis. Computation by itself cannot furnish this structure and there are many problems involving errors of statistical nature which are better considered by means of the theoretical structure given here than by mass computation.

1.3. The α, β and λ Errors. We suppose that the system to be solved is given to us in the form of a system of n first order differential equations

$$F_i(\dot{x}_1, \cdots, \dot{x}_n, x_1, \cdots, x_n, t) = 0, \qquad i = 1, \cdots, n. \tag{1.1}$$

The precise conditions on the F_i will be stated in §2.1 below but we may indicate the situation assumed as follows. Equations (1.1) are to be considered on a region R in $n + 1$ dimensional space of the variables $x_1, \cdots, x_n, t$. We suppose that R can be divided into subregions on each of which we can solve for the $\dot{x}_i$ in terms of $x_1, \cdots, x_n$ and t:

$$\dot{x}_i = f_i(x_1, \cdots, x_n, t), \qquad i = 1, \cdots, n \tag{1.2}$$

where f_i is analytic in $x_1, \cdots, x_n$ and continuous in $x_1, \cdots, x_n, t$ for the subregion. A rule is given for continuing a solution from one subregion to another when a boundary point is reached.

A system such as Eqs. (1.1) will be realized on the machine in the form

$$G_i = 0, \qquad i = 1, \cdots, n \tag{1.3}$$

where the G_i's differ from the F_i's due to our inability to accurately realize the original system on the machine.

The variations from F_i to G_i may affect the order of the system, i.e., they may introduce higher derivatives of the x_j. Errors of this type we call λ errors. Variations which do not affect the order will be called α errors. For each type of error we introduce parameters, λ and $\alpha_1, \cdots, \alpha_N$, respectively into the system of

Eq. (1.3) so that if all the parameters are zero, Eq. (1.3) becomes Eq. (1.1) while if they assume certain other non-zero values we obtain Eq. (1.3) as realized on the machine.

In addition to α and λ errors the machine solution may be affected by another type of error which we shall term β errors. β errors arise in the course of a machine computation as the result of instantaneous disturbances of the solution. These are disturbances which appear in the solution but which do not appear in the differential equations of motion, Eq. (1.3), for the machine. For example, suppose we have a solution of the machine equations and at a certain point in the solution process the variable x_i was arbitrarily disturbed and instantaneously changed by an amount β. After the change, the machine would continue on a solution of the original system of equations, as it did before. But the effect of this instantaneous perturbation is to jar the machine from one solution to another. We may also consider as β errors the errors made in setting up the initial conditions of a given solution. Furthermore, in general, when a machine passes from one region of analyticity to another in the course of the solution, it will not follow out precisely the rule given for this change. This discrepancy can be described in terms of a β error. Still another example is given by integrating amplifier output noise. Such noise can be described in terms of a "shot" effect, that is, in terms of disturbances of the above sort in the solution which occurs in time according to a certain probability distribution and whose magnitude is governed by another probability distribution.

1.4. The α, β Error Theory and Sensitivity. Our plan of attack on the general error theory involves first a discussion of the case in which λ is considered fixed. The results obtained in this case are used to show that the general case in which λ errors appear can be referred to a problem involving linear systems with constant coefficients. The latter problem can be solved. (Cf. Chap. 2, also, Macnee [28].)

We begin by considering a situation in which we have only α and β errors, i.e., no λ errors. Suppose, then, that we have N α errors $\alpha_1, \cdots, \alpha_N$ and M β errors, $\beta_1, \cdots, \beta_M$. In this case it is possible to apply a generalization of the usual existence theorems for systems of ordinary differential equations which depend on parameters. As a consequence one can show that the functions given by the machine

$$x_i = x_i(t, \alpha_1, \cdots, \alpha_N, \beta_1, \cdots, \beta_M) \tag{1.4}$$

depend analytically on the parameters $\alpha_1, \cdots, \alpha_N, \beta_1, \cdots, \beta_M$. (We suppose the initial conditions are fixed. Any error in realizing them will be a β error.)

Equation (1.4) reduces to the correct solution when we set the α's and β's equal to zero and corresponds to the inaccurate machine solution for certain values of these parameters. We can use the analytic dependence of our solution in this case to express the solutions x_i as power series in the α's and β's.

$$x_j = x_j(t, 0, \cdots, 0, 0, \cdots, 0) + \sum_k \frac{\partial x_j}{\partial \alpha_k} \alpha_k + \sum_l \frac{\partial x_j}{\partial \beta_l} \beta_l$$
$$+ \frac{1}{2!}\left[\sum_{k_1,k_2} \frac{\partial^2 x_j}{\partial \alpha_{k_1} \partial \alpha_{k_2}} \alpha_{k_1}\alpha_{k_2} + \sum_{k_1,l_1} \frac{\partial^2 x_j}{\partial \alpha_{k_1} \partial \beta_{l_1}} \alpha_{k_1}\beta_{l_1}\right. \tag{1.5}$$
$$\left. + \sum_{l_1,l_2} \frac{\partial^2 x_j}{\partial \beta_{l_1} \partial \beta_{l_2}} \beta_{l_1}\beta_{l_2}\right] + \cdots.$$

We must, of course, assume that the errors α and β are not so large that the series diverges. One might even advance the argument that if our error analysis requires that we go further than, say, the third degree terms in the above expansion, then the solution is so far off as to be of little value. However, theoretically, as long as the series converges we may use the above expression to obtain estimates for the effect of errors. Thus, we see that our error analysis can be reduced, in the case considered, to a study of the partial derivatives,

$$\frac{\partial x_j}{\partial \alpha}, \quad \frac{\partial x_j}{\partial \beta}, \quad \frac{\partial^2 x_j}{\partial \alpha^2}, \quad \frac{\partial^2 x_j}{\partial \alpha \partial \beta}, \quad \frac{\partial^2 x_j}{\partial \beta^2}, \text{ etc.}$$

Now if we let w_j stand for any of these, considered as functions of t then by substituting in Equations (1.3) and taking the appropriate partials with respect to α or β, and then letting $\alpha = 0 = \beta$, we find that the w_j for $j = 1, \cdots, n$ satisfy a system of differential equations

$$\sum_j \frac{\partial F_i}{\partial \dot{x}_j} \dot{w}_j + \sum_k \frac{\partial F_i}{\partial x_k} w_k + T_i = 0 \tag{1.6}$$

where the T_i involve only partial derivatives of the x_k relative to α and β of lower order. Thus the w_j's can be obtained by an inductive process.

For all orders the homogeneous part of the system (1.6) is the same and it is known that the solutions of Eqs. (1.6) can be readily obtained if we solve the corresponding homogeneous system.

$$\sum_j \frac{\partial F_i}{\partial \dot{x}_j} \dot{w}_j + \sum_k \frac{\partial F_i}{\partial x_k} w_k = 0 \tag{1.7}$$

Consequently, the question of sensitivity to error of our solutions can be referred back to the study of this system of linear homogeneous differential equations. We will refer to Eqs. (1.7) as the *sensitivity equations*. Thus, in the case in which there are no λ errors, we have succeeded in referring our problem to the solution of a system of linear differential equations without any unjustified linearization of the given problem.

1.5. The λ Errors. We now return to a discussion of the λ errors. For the moment we ignore α and β errors. Furthermore, suppose here for simplicity that the order is raised at most one, and also for simplicity in exposition, that a rise occurs in every equation. We then can write the disturbed system of equa-

tions in the form

$$G_i(\lambda\ddot{x}_1, \cdots, \lambda\ddot{x}_n, \dot{x}_1, \cdots, \dot{x}_n, x_1, \cdots, x_n, t) = 0. \tag{1.8}$$

Equation (1.8) is now the equation of motion for the machine. A parameter λ is introduced as a multiplier of the second derivatives so that one can make them all vanish from the equations simultaneously. One thinks of λ as small, while the dependence on $\lambda\ddot{x}_j$ is to be of reasonable size. Let $x_1', \cdots, x_n'$ be the solution of Eq. (1.1) corresponding to the given initial conditions and let $x_1, \cdots, x_n$ be the machine solution which satisfies Eq. (1.8). Suppose we define u_j by the equations

$$x_j = x_j' + u_j \tag{1.9}$$

and substitute this into Eq. (1.8).

We then obtain

$$G_i(\lambda\ddot{x}_1' + \lambda\ddot{u}_1, \cdots, \lambda\ddot{x}_n' + \lambda\ddot{u}_n, \dot{x}_1' + \dot{u}_1, \cdots, \dot{x}_n' + \dot{u}_n, x_1' + u_1, \cdots, x_n' + u_n, t) = 0. \tag{1.10}$$

We can expand this in powers of $\lambda\ddot{u}_j$, $\dot{u}_j$ and u_j, $j = 1, \cdots, n$.

$$G_i = G_i(\lambda\ddot{x}_1', \cdots, \lambda\ddot{x}_n', \dot{x}_1', \cdots, \dot{x}_n', x_1', \cdots, x_n', t) + \sum_j \frac{\partial G_i}{\partial \lambda\ddot{x}_j}\lambda\ddot{u}_j + \sum_j \frac{\partial G_i}{\partial \dot{x}_j}\dot{u}_j + \sum_j \frac{\partial G_i}{\partial x_j}u_j + R_i \tag{1.11}$$

where R_i, of course, depends on higher powers of $\lambda\ddot{u}_j$, $\dot{u}_j$, and u_j. The partial derivatives $\partial G_i/\partial(\lambda\ddot{x}_j)$, $\partial G_i/\partial\dot{x}_j$, and $\partial G_i/\partial x_j$ now depend only on the variable t and the parameter λ since the other variables $\ddot{x}_i$, $\dot{x}_i$ and x_i have been replaced by the functions $\ddot{x}_i'$, $\dot{x}_i'$, and x_i' of t. These partials are supposed to be continuous in the variable t. Consequently, if we divide our region into small enough pieces we can suppose that on each piece these partials are constants. Thus, we write

$$G_i = G_i(\lambda\ddot{x}_1', \cdots, \lambda\ddot{x}_n', \dot{x}_1', \cdots, \dot{x}_n', x_1', \cdots, x_n', t) + \sum_j A_{ij}\lambda\ddot{u}_j + \sum_j B_{ij}\dot{u}_j + \sum_j C_{ij}u_j + \sum_j \left(\frac{\partial G_i}{\partial \lambda\ddot{x}_k} - A_{ij}\right)\lambda\ddot{u}_j + \sum_j \left(\frac{\partial G_i}{\partial \dot{x}_j} - B_{ij}\right)\dot{u}_j + \sum_j \left(\frac{\partial G_i}{\partial x_j} - C_{ij}\right)u_j + R_i. \tag{1.12}$$

Now, this equation differs from the equation with constant coefficients on the subregions by small terms. So, we can introduce into this expression a parameter η

$$G_i(\cdots, \cdots, t, \eta) = \eta G_i(\lambda\ddot{x}_1', \cdots, \dot{x}_1', \cdots, x_1', \cdots, t) + \sum_j A_{ij}\lambda\ddot{u}_j + \sum_j B_{ij}\dot{u}_j + \sum_j C_{ij}u_j + \eta\left[\sum_j \left(\frac{\partial G_i}{\partial \lambda\ddot{x}_j} - A_{ij}\right)\lambda\ddot{u}_j + \sum_j \left(\frac{\partial G_i}{\partial \dot{x}_j} - B_{ij}\right)\dot{u}_j + \sum_j \left(\frac{\partial G_i}{\partial x_j} - C_{ij}\right)u_j\right] + \eta R_i. \tag{1.13}$$

We now compare Eq. (1.8) with the system

$$\sum_j A_{ij}\lambda\ddot{u}_j + \sum_j B_{ij}\dot{u}_j + \sum_j C_{ij}u_j = 0 \tag{1.14}$$

where, of course, the A_{ij}, B_{ij}, and C_{ij} vary with the subdivision. We can pass from Eqs. (1.14) to our original problem Eqs. (1.8) by introducing the parameter η as in Eq. (1.13). Consequently, the analytic behavior of the solution of Eq. (1.10) must be the same as that of Eq. (1.14). This means, then, that we can use our α techniques to reduce the study of time delay errors to the linear case. [For $\eta = 0$, Eq. (1.13) reduces to Eq. (1.14) and for $\eta = 1$, Eq. (1.13) reduces to Eq. (1.11).]

We can regard Eq. (1.14) as the machine equation for the system

$$\sum_j B_{ij}\dot{u}_j + \sum_j C_{ij}u_j = 0 \tag{1.15}$$

and pass from Eq. (1.14) to Eq. (1.8) by supposing u to depend on a parameter η as in Eq. (1.13).

The above is the basic principle upon which the theory of this paper rests. The formulae above have been simplified by assuming a uniform rise in the order. This assumption is not made in Chapter II.

The discussion below proceeds to consider the effect of λ errors on the system of Eq. (1.15). One can show that, in general, the roots of the indicial equation of Eq. (1.14), which would be a polynomial of degree $2n$, can be divided into two sets $\mu_1, \cdots, \mu_n$ and $\nu_1/\lambda, \cdots, \nu_n/\lambda$. Both the $\mu_1, \cdots, \mu_n$ and $\nu_1, \cdots, \nu_n$ are, in general, analytic in λ at $\lambda = 0$. The corresponding solutions of Eq. (1.14) have exponential factors

$$e^{\mu_i t} \quad \text{and} \quad e^{(\nu_i/\lambda)t}.$$

The μ_i terms correspond to what one would normally consider to be the long range effect of errors. On the other hand, the ν_i/λ exponentials are not analytic in λ at $\lambda = 0$ and either destroy the solution completely when the real part of ν_i is positive or if the real part of ν_i is negative they become very small in a brief t interval. The total error u when λ is present can be expressed in two sets of terms, one of which has the above mentioned μ properties, the other the above mentioned ν properties.

In §2.8, the combined effect of α, β, and λ errors are considered. Let us consider again the machine Eq. (1.8) with however α errors added and with β errors permitted in the solution. We may again introduce the parameter η to yield the equivalent of Eqs. (1.13). η is an α parameter. Thus we may obtain for the machine solution an expansion about the point where $\alpha = 0$, $\beta = 0$ in the form

$$x(t, \lambda, \alpha, \beta) = x(t, \lambda, 0, 0) + \sum_\gamma x_\gamma(t, \lambda, 0, 0)\gamma + \cdots. \tag{1.16}$$

Here, $x(t, \lambda, 0, 0)$ is the solution obtained under the assumption of no α or β errors, while the partials of x with respect to the various γ's are obtained by an argument similar to that for the α, β error, using however Eqs. (1.14) as sensitivity equations.

2.1. The Given Mathematical Problem. The purpose of the present section is to describe precisely the types of problems to which the present error analysis is applicable. This description is given at the end of this section.

A differential analyzer presents the solution of a system of differential equations,

$$F_j(\dot{x}_1, \cdots, \dot{x}_n, x_1, \cdots, x_n, t) = 0, \qquad j = 1, \cdots, n \tag{2.1}$$

where t is the independent variable, $x_1, \cdots, x_n$ are n unknown functions of t and $\dot{x}_i$ stands for the derivatives of x_i with respect to t. In addition to Eq. (2.1), initial values for the n variables $x_1, \cdots, x_n$ at t_0 are given. It is always possible to express a differential equation problem in a form in which only first derivatives appear and we will suppose that this has been done.

In many applications of interest one is not justified in considering equations $F_j = 0$ as given by a single analytic expression defined by a region in $n + 1$ dimensional space of variables $x_1, \cdots, x_n, t$. Instead, it is frequently necessary to consider the region of interest R as broken up into smaller regions on each of which Eq. (2.1) can be solved for the $\dot{x}_i$,

$$\dot{x}_i = f_i(x_1, \cdots, x_n, t),$$

where f_i is analytic on each subregion in $x_1, \cdots, x_n$ for t fixed and is Riemann integrable in t when continuous functions $x_1(t), \cdots, x_n(t)$ are substituted for $x_1, \cdots, x_n$. The boundaries between these regions are analytic manifolds of not more than n dimensions. We are justified in including a boundary point in one of the subregions if the functions f_i satisfy the above mentioned criterion at such a boundary point.

We shall suppose that the operation of the machine indicated the existence of a solution. In general we will be interested in solutions which extend through one or more of the boundary surfaces between regions of analyticity. If a point on such a solution is on the boundary between two regions and the f's associated with both regions are analytic at this boundary point, then the continuation of the solution through the boundary point is determined by the uniqueness theorem for ordinary differential equations. Furthermore, the analytic character of the dependence of the solution on its initial conditions remains after passing a boundary point in this fashion. On the other hand, we require in our discussion below this analytic dependence on the initial conditions and we shall take this as our hypothesis rather than specifications concerning the behavior of f on the boundary. We now give a precise statement of the hypotheses needed in the discussion below.

Let

$$F_j(\dot{x}, x, t) = 0, \qquad j = 1, \cdots, n \tag{2.1}$$

denote a system of n first order ordinary differential equations in which t is the independent variable, x stands for n dependent variables $x_1, \cdots, x_n$ *and* $\dot{x}$ *stands for the n derivatives* $\dot{x}_1, \cdots, \dot{x}_n$ *of* $x_1, \cdots, x_n$ *with respect to t. A solution of this system consisting of n functions* $x_1(t), \cdots, x_n(t)$ *is desired which at* $t = t_0$ *assumes specified values* $x_{1,0}, \cdots, x_{n,0}$, *e.g.* $x_i(t_0) = x_{i,0}$.

The Eqs. (2.1) *are to be considered on a region* R *in the* $n + 1$ *dimensional space of the variables* $x_1, \cdots, x_n, t$. *This region* R *is subdivided into subregions* $R_1, R_2, \cdots$ *which do not have interior points in common. On* R_k *it is possible to solve Eqs.* (2.1) *for* $\dot{x}_1, \cdots, \dot{x}_n$ *in terms of* $x_1, \cdots, x_n, t$,

$$\dot{x}_i = f_i^k(x_1, \cdots, x_n, t) \tag{2.2}$$

where:

(*A*) *If* $x_1', \cdots, x_n', t'$ *is a point of* R_k *then for* t' *fixed,* $f_i^k(x_1, \cdots, x_n, t')$ *is analytic in* $x_1, \cdots, x_n$ *at* $x_i', \cdots, x_n'$ *and:*

(*B*) *If* $x_1(t), \cdots, x_n(t)$ *are* n *continuous functions of* t *defined for an interval* $a \leqq t \leqq b$ *and such that for every* t *in this interval, the point* $x_1(t), \cdots, x_n(t), t$ *is in* R_k, *then* $f_i^k(x_1(t), \cdots, x_n(t), t)$ *is a Riemann integrable function of* t *for a* $a \leqq t \leqq b$.

Let B_{kl} *denote the intersection of the boundary of* R_k *and* R_l. We *suppose that if* B_{kl} *is not null it can be specified by a finite number of analytic equations and inequalities. If the addition of* B_{kl} *to* R_k *does not destroy the properties* (*A*) *and* (*B*) *above, we shall consider* B_{kl} *to be in* R_k. *Similarly for* R_l. *Thus,* R_k *and* R_l *may have boundary points in common.*

Let $x_{1,0}, \cdots, x_{n,0}, t_0$ *be an interior point of a subregion which we suppose has been denoted* R_1. *Since* f^1 *satisfies conditions* (*A*) *and* (*B*) *on* R_1, *there is a unique solution,* $x_1(t), \cdots, x_n(t)$, *of Eq.* (2.2) *which passes through* $x_{1,0}, \cdots, x_{n,0}, t_0$. *Suppose that at* $t \to t_1$, $t_0 \leqq t < t_1$, $x_j(t) \to x_{j1}$ *where* $x_{11}, \cdots, x_{n1}, t_1$ *is a point of* B_{12}, *but not of* B_{1j} *for* $j \neq 2$. *A continuation rule is a rule which associates with every such solution defined for* $t_0 \leqq t \leqq t_1$ *a unique solution of Eq.* (2.2) *defined for a* t *interval with lower endpoint* t_1 *and with* $x_1(t), \cdots, x_n(t)$, t *in* R_2 *for* $t > t_1$. *We suppose that such a continuation rule exists for every non-null boundary* B_{jk}.

If B_{12} belongs to both R_1 and R_2 then there is only one continuation rule possible. This is a consequence of the usual uniqueness theorem.

We shall say that the above mentioned solution penetrates the boundary analytically, if we can find a $\delta > 0$ *and a* $\delta' > 0$ *such that if* $x_j' = x_j(t_1 - \delta)$, *then there exists a neighborhood of* $x_1', \cdots, x_n'$ *such that:*

(1) *Each point* $x_1, \cdots, x_n$ *of this neighborhood is such that* $x_1, \cdots, x_n, t_1 - \delta$ *is in* R_1,

(2) *Each point,* $x_1, \cdots, x_n$ *of this neighborhood determines a solution* $x_1(t), \cdots, x_n(t)$, t *which can be continued by the continuation rule through* B_{12} *into* R_2,

(3) *The values of the continuation* $x_1(t_1 + \delta'), \cdots, x_n(t_1 + \delta')$, $t_1 + \delta'$ *are in* R_2 *for every such point and*

(4) *Each such* $x_j(t + \delta')$ *is an analytic function of* $x_1, \cdots, x_n$ *in the neighborhood.*

Again, it may be stated that if B_{12} belongs to both R_1 and R_2, then every continued solution penetrates the boundary analytically provided the point of penetration is not a boundary point of a region other than R_1 and R_2. If B_{12} is in R_2, the continuation $x_1(t), \cdots, x_n(t)$, t, $t \geqq t_1$ may remain in B_{12}.

2.2. The Machine Realization. The effort to obtain a solution of the system of differential equations, Eq. (2.1), with given initial conditions $x_{1,0}, \cdots, x_{n,0}, t_0$ is beset with two major difficulties. In the first place, the system of equations $F_i = 0$ cannot be accurately realized on the machine. Instead, we actually realize another system, $G_i = 0$. We suppose that the difference between the two systems can be described as follows. The G_i's are dependent on certain parameters, $\lambda, \alpha_1, \cdots, \alpha_N$ such that $\lambda = 0, \alpha_j = 0$, the equations $G_i = 0$ reduce to the system $F_i = 0$. The order of the system $G_i = 0$, may be higher than the original system. In addition, however, the process of solution on the machine may also be subject to perturbations during the course of the solution.

For simplicity in our present discussion, we will assume that the order of the system $G_i = 0$ has increased only by the introduction of second derivatives $\ddot{x}_1, \cdots, \ddot{x}_n$, i.e. the system realized on the machine will be in the form

$$G_i(\lambda\ddot{x}_1, \cdots, \lambda\ddot{x}_n, \dot{x}_1, \cdots, \dot{x}_n, x_1, \cdots, x_n, t, \alpha_1, \cdots, \alpha_N) = 0, \quad i = 1, \cdots, n \tag{2.3}$$

or as we prefer to write it

$$G(\lambda\ddot{x}, \dot{x}, t, \alpha) = 0.$$

The parameter λ has been introduced for the purpose of convenience. It is convenient to think of λ as small and that $\partial G/\partial(\lambda\ddot{x})$ is of normal size. The assumption that the order only increases by one is made in order to obtain reasonably compact formulas. The arguments will generalize to the cases in which higher derivatives appear. For the moment, we will suppose that the α's are held fixed and for a time reference will not be made to them.

We must first, however, specify the extent to which Eqs. (2.3) are actually of the second degree. Suppose the Eqs. (2.3) imply exactly $n - r$ independent functional relations

$$G_j(\dot{x}_1, \cdots, \dot{x}_n, x_1, \cdots, x_n, t) = 0, \qquad j = r + 1, \cdots, n \tag{2.4}$$

which do not involve the second derivatives. Now one can readily see that if Eq. (2.3) is such that we can explicitly solve for s of the expressions $\lambda\ddot{x}_1, \cdots, \lambda\ddot{x}_s$ in terms of $\lambda\ddot{x}_{s+1}, \cdots, \lambda\ddot{x}_n, \dot{x}_1, \cdots, \dot{x}_n, x_1, \cdots, x_n, t$, then when we substitute these in the remaining $n - s$ equations of Eq. (2.3) either the $\lambda\ddot{x}_{s+1}, \cdots, \lambda\ddot{x}_n$ appear and we can solve for another $\lambda\ddot{x}$ or they disappear from the remaining equations and we have $n - s$ relations of the type Eq. (2.3). Since there are exactly $n - r$ relations in Eq. (2.4), we can carry out the above elimination process for $s = 1, \cdots, r$. Suppose then that the first r equations

$$G_i(\lambda\ddot{x}_1, \cdots, \lambda\ddot{x}_n, \dot{x}_1, \cdots, \dot{x}_n, x_1, \cdots, x_n, t) = 0, \qquad i = 1, \cdots, r \tag{2.5}$$

can be used for solving for $\lambda\ddot{x}_1, \cdots, \lambda\ddot{x}_r$ and that Eqs. (2.5) and (2.4) form a system equivalent to Eq. (2.3).

Now suppose the system of Eq. (2.4) is adequate to specify $\dot{x}_{r+1}, \cdots, \dot{x}_n$ in terms of $\dot{x}_1, \cdots, \dot{x}_r, x_1, \cdots, x_n, t$. This means that a certain Jacobian is not

zero. Thus, if we differentiate Eq. (2.4) with respect to t,

$$\sum_i \frac{\partial G_j}{\partial \dot{x}_i} \ddot{x}_j + \sum_i \frac{\partial G_j}{\partial x_i} \dot{x}_i + \frac{\partial G_j}{\partial t} = 0, \qquad j = r+1, \cdots, n, \tag{2.6}$$

and these equations can be used to eliminate $\lambda\ddot{x}_{r+1}, \cdots, \lambda\ddot{x}_n$ from Eqs. (2.5). In order to present the following discussion with a reasonable economy we will suppose that this latter elimination has been carried out.

Hypothesis 2.1. *The system of Eq.* (2.3) *can be written*

$$\begin{aligned} G_i'(\lambda\ddot{x}_1, \cdots, \lambda\ddot{x}_r, \dot{x}_1, \cdots, \dot{x}_n, x_1, \cdots, x_n, t, \lambda, \alpha) &= 0 \qquad i = 1, \cdots, r \\ G_i'(\dot{x}_1, \cdots, \dot{x}_n, x_1, \cdots, x_n, t, \alpha) &= 0 \qquad i = r+1, \cdots, n \end{aligned} \tag{2.7}$$

where the first r *equations can be used to solve for* $\lambda\ddot{x}_1, \cdots, \lambda\ddot{x}_r$ *in terms of* $\dot{x}_1, \cdots, \dot{x}_n, x_1, \cdots, x_n, t$ *and the remaining equations can be used to solve for* $\dot{x}_{r+1}, \cdots, \dot{x}_n$ *in terms of* $\dot{x}_1, \cdots, \dot{x}_r, x_1, \cdots, x_n$. *Hypothesis* 2.1 *is to be understood as stating that certain Jacobians are not zero plus the existence of certain values satisfying the equations. We suppose then that the equations of motion of the machine can be written in the form of Eq.* (2.7) *and that at* $\lambda = 0$, $\alpha = 0$ *these become equivalent to Eq.* (2.1).

In the following discussion r will be considered to be fixed for the given problem. This is purely for economy in notation. One can generalize the discussion to a case in which r is different on the different regions of analyticity of the problem.

The equations (2.7) contain various parameters α corresponding to errors in the machine realization which do not raise the order of the system. However in addition to the errors which appear in the realization of the system of differential equations, the machine solution is subject to perturbations during the course of the run. The variables $\dot{x}_1, \cdots, \dot{x}_r, x_1, \cdots, x_n$ may contain discontinuities corresponding to a jump from one solution to another. These will be referred to as β errors. They may arise accidently or as errors in entering the initial conditions or as errors in realizing the continuation conditions when a solution passes from one region of non-analyticity to another or as output noise.

Hypothesis 2.2. *We assume that the motion of the machine is described by* n *functions* $x_1(t), \cdots, x_n(t)$ *of the time* t *which satisfy a system of differential equations, Eq.* (2.3), *dependent on* $N+1$ *parameters* α_j *and* λ. *When every* $\alpha = 0$, *and* $\lambda = 0$ *the system of Eq.* (2.3) *reduces to a system of equations* $F_i = 0$, [*Eq.* (2.1)].

The Eqs. (2.3) *are to be considered on a region* R *in* $r + n + N + 1$ *dimensional space of the variables* $\dot{x}$, x, α, t. *This region is to be subdivided into subregions* $R_1, R_2, \cdots$, *which do not have interior points in common. On each* R_k *Hypothesis* 2.1 *may be applied to solve Eqs.* (2.7) *for* $\lambda\ddot{x}_1, \cdots, \lambda\ddot{x}_r, \dot{x}_r, \dot{x}_{r+1}, \cdots, \dot{x}_n$ *in terms of the other variables.*

$$\begin{aligned} \lambda\ddot{x}_i &= g_i^k(\dot{x}_1, \cdots, \dot{x}_r, x_1, \cdots, x_n, t, \alpha_1, \cdots, \alpha_N) \qquad i = 1, \cdots, r \\ \ddot{x}_i &= g_i^k(\dot{x}_1, \cdots, \dot{x}_r, x_1, \cdots, x_n, t, \alpha_1, \cdots, \alpha_N) \qquad i = r+1, \cdots, n \end{aligned} \tag{2.8}$$

where

(A) If $\dot{x}', x', \alpha', t'$ is a point of R_k then for t' fixed, g_i^k is analytic in $\dot{x}_1, \cdots, \dot{x}_r, x_1, \cdots, x_n, \alpha_1, \cdots, \alpha_N$ at $\dot{x}_1', \cdots, \dot{x}_r', x_1', \cdots, x_n', \alpha_1', \cdots, \alpha_N'$ and

(B) If $\dot{x}_1(t), \cdots, \dot{x}_r(t), x_1(t), \cdots, x_n(t)$ are $n + r$ continuous functions of t defined for an interval $a \leqq t \leqq b$ and $\alpha_1, \cdots, \alpha_N$ are fixed values of α such that for every t in the given interval, $\dot{x}_1(t), \cdots, \dot{x}_r(t), x_1(t), \cdots, x_n(t), \alpha_1, \cdots, \alpha_N, t$ is in R_k, then $g_i^k(\dot{x}_1(t), \cdots, \dot{x}_r(t), x_1(t), \cdots, x_n(t), t, \alpha_1, \cdots, \alpha_N)$ are Riemann integrable in t for $a \leqq t \leqq b$.

At $\lambda = 0$, the Eq. (2.3) can be solved for $\dot{x}_1, \cdots, \dot{x}_n$, to yield

$$\dot{x}_i = g_i'(x_1, \cdots, x_n, t, \alpha_1, \cdots, \alpha_n)$$

where the g_i' satisfy conditions A', B' analogous to A, B above with reference to $\dot{x}_1, \cdots, \dot{x}_r$ omitted.

The boundary B_{kl} between the regions R_k and R_l is subject to a description similar to that for B_{kl} in the previous section with the addition of reference to the $\dot{x}_1, \cdots, \dot{x}_r$ and for a fixed set of values $\alpha_1, \cdots, \alpha_N$, the definition of a continuation rule given in the previous section applies here.

Suppose that for a fixed set of values, $\alpha_1', \cdots, \alpha_N'$ of the parameters, a solution of Eq. (2.3) is given with initial conditions $\dot{x}_{1,0}, \cdots, \dot{x}_{r,0}, x_{1,0}, \cdots, x_{n,0}, t_0$ such that $\dot{x}_{1,0}, \cdots, \dot{x}_{r,0}, x_{1,0}, \cdots, x_{n,0}, \alpha_1', \cdots, \alpha_N', t'$ is in R_1. Suppose that as $t \to t_1$, $t_0 \leqq t < t_1$, $\dot{x}_j(t) \to \dot{x}_{j1}$, $x_j(t) \to x_{j1}$ where $\dot{x}_{11}, \cdots, \dot{x}_{r1}, x_{11}, \cdots, x_{n1}, \alpha_1', \cdots, \alpha_N', t_1$ is a point of B_{12} but not of B_{1j} for $j \neq 2$. For $t > t_1$ let $x_j(t)$ denote the continuation of this solution into the region R_2. We say that this solution penetrated the boundary analytically, if we can find a $\delta > 0$ and a $\delta' > 0$ such that if $\dot{x}_j' = \dot{x}_j(t_1 - \delta)$, $x_j' = x_j(t_1 - \delta)$, then there exists a neighborhood of $\dot{x}_1', \cdots, \dot{x}_r', x_1', \cdots, x_n', \alpha_1', \cdots, \alpha_N'$ such that:

(1) Each point of this neighborhood $\dot{x}_1, \cdots, \dot{x}_r, x_1, \cdots, x_n, \alpha_1, \cdots, \alpha_N$ is such that $\dot{x}_1, \cdots, \dot{x}_r, x_1, \cdots, x_n, \alpha_1, \cdots, \alpha_N, t_1 - \delta$ is in R_1,

(2) Each point $\dot{x}_1, \cdots, \dot{x}_r, x_1, \cdots, x_n, \alpha_1, \cdots, \alpha_N$ determines a solution which can be continued by the continuation rule through B_{12} into R_2,

(3) The values of the continuation at $t + \delta'$ determine a point $\dot{x}_1(t + \delta'), \cdots, \dot{x}_r(t + \delta'), x_1(t + \delta'), \cdots, x_n(t + \delta'), \alpha_1, \cdots, \alpha_N, t + \delta'$ which is in R_2 for every such point in the neighborhood and,

(4) Each such $x_j(t + \delta')$ is analytic in $\dot{x}_1, \cdots, \dot{x}_r, x_1, \cdots, x_n, \alpha_1, \cdots, \alpha_N$ in this neighborhood.

Let the initial conditions $\dot{x}_{1,0}, \cdots, \dot{x}_{r,0}, x_{1,0}, \cdots, x_{n,0}$ at t_0 and $\alpha_1, \cdots, \alpha_N$ be fixed. A set of functions $x_1(t), \cdots, x_n(t)$ will be said to describe a perturbed motion of the machine if a set of values $t_1, \cdots, t_M$ of t are given such that:

(i) For $t_i < t < t_{i+1}$, $x_1(t), \cdots, x_n(t)$ is a solution of Eqs. (2.3) except possibly for a finite number of values of t where this solution analytically penetrates a boundary and:

(ii) At each t_j we have parameters $\dot{\beta}_{ij}$, β_{ij} such that $\dot{x}_i(t_j+) = \dot{x}_i(t_j) + \dot{\beta}_{ij}$, $x_i(t_j+) = x_i(t_j) + \beta_{ij}$. [The form of the last equation permits us to assume that $x_i(t_0) = x_{i,0}$, even when t_0 is a point of perturbation, t_1, i.e. $x_i(t_0+) = x_{i,0} + \beta_{i1}$.]

It is convenient mathematically to permit "perturbed motions" in which all the β's are zero.

Hypothesis 2.3. *Let a set of initial conditions* $\dot{x}_{1,0}, \cdots, \dot{x}_{r,0}, x_{1,0}, \cdots, x_{n,0}, t_0$ *be given. The Eq.* (2.3) *and the value* λ *are such that for* $\alpha = 0$ *and for all* $\hat{\beta}_{i,j}, \beta_{i,j} = 0$ *there exists a perturbed motion perturbed at the specified values of* t, *defined for a closed* t *interval, with lower end point* t_0 *and which meets at most a finite number of boundaries and penetrates them analytically.*

THEOREM 1. *Under Hypotheses* 2.1, 2.2, *and* 2.3, *there exists a constant* $\delta > 0$, *such that for* $|\dot{x}_i' - \dot{x}_{i,0}| < \delta$, $|x_i' - x_{i,0}| < \delta$, $|\alpha_i| < \delta$, $|\hat{\beta}_{ij}| < \delta$, $|\beta_{ij}| < \delta$, *there exists perturbed motions in the sense of the above for the* t *interval of Hypothesis* 2.3, *with initial conditions* $\dot{x}_1', \cdots, \dot{x}_r', x_1', \cdots, x_n'$, *satisfying Eqs.* (2.3) *for the given values of* λ *and* α *and having the given values of* β *as perturbations at the specified points. For each value of* t *in the closed interval, the functions* $x_i(t)$ *which describe the perturbed functions are analytic functions of* $\dot{x}_1', \cdots, \dot{x}_r', x_1', \cdots, x_n'$, *the* α's *and* β's *on the neighborhood described above in terms of* δ.

The Hypotheses given above have been set up so that at each point $t, \dot{x}_1, \cdots, \dot{x}_r, x_1, \cdots, x_n$ on the "perturbed motion" of Hypothesis 2.3 it is possible to find a closed t interval, having the given point as an interior point for which the theorem holds. If the end points of these closed intervals are ignored, we can apply the Heine-Borel theorem to obtain a finite number of open intervals, which cover the given interval of the theorem in an overlapping manner. The analyticity in the initial conditions permit us to combine these functions when overlap is present so that the theorem holds for the full given interval of Hypothesis 2.3.

The analyticity in the initial conditions is used only in this proof. We will require only the analyticity in α and β in the future. In the proof of the theorem, "analyticity in the $\dot{x}_1, \cdots, \dot{x}_r, x_1, \cdots, x_n$" is used to permit the application of standard existence theorems. If necessary, one may replace "analyticity" throughout by the condition "having continuous kth order partial derivatives, $k \geqq 1$."

2.3. The Linearization. Now let $x_1', \cdots, x_n'$ be a given solution of Eq. (2.1) corresponding to a prescribed set of initial conditions $x_{1,0}', \cdots, x_{n,0}'$. We shall write the solution of Eq. (2.7) obtained when one tries to solve Eq. (2.1) on the machine in the form

$$x_i = x_i' + u_i. \tag{2.9}$$

Clearly, the initial values of $x_1, \cdots, x_n$ and $\dot{x}_1, \cdots, \dot{x}_r$ are determined by corresponding initial values of $u_1, \cdots, u_n, \dot{u}_1, \cdots, \dot{u}_r$. We may consider Eq. (2.7) then as determining a system of equations for the u_i

$$G_i'(\lambda'\dot{x}_1' + \lambda\dot{u}_1, \cdots, \lambda'\dot{x}_r' + \lambda\dot{u}_r, \dot{x}_1' + \dot{u}_1, \cdots, x_n' + u_n, t) = 0, \quad i = 1, \cdots, r \tag{2.10}$$

$$G_i'(\dot{x}_1' + \dot{u}_1, \cdots, \dot{x}_n' + \dot{u}_n, x_1' + u_1, \cdots, x_n' + u_n, t) = 0, \quad i = r+1, \cdots, n.$$

We distinguish between λ as a coefficient of $\dot{u}$ and λ' as a coefficient of $\dot{x}'$. The latter λ' can be conveniently regarded as another α parameter.

Normally one is justified in regarding the initial values of the u_i's as small. This means that the x_i's have been entered into the machine with approximately the correct values. However, we do not have any assurance that the values of the $\dot{u}_j$'s are small. We shall give a discussion which will clearly have a t interval of validity if $\dot{u}$ is small and this will certainly happen if the initial value of $\dot{u}$ is small. On the other hand, in general this discussion will have an adequate range of validity but this cannot be presupposed if $\dot{u}_i$ is arbitrary. We shall make this matter precise later.

We may write Eq. (2.10) in the form

$$G_i'(\lambda'\dot{x}', \dot{x}', x', t) + \sum_j \frac{\partial G_i'}{\partial(\lambda\dot{x}_j)}(\lambda\dot{u}_j) + \sum_j \frac{\partial G_i'}{\partial \dot{x}_j}\dot{u}_j + \sum_j \frac{\partial G_i'}{\partial x}u + R_i = 0, \qquad i = 1, \cdots, r \tag{2.11}$$

$$G_i'(\dot{x}', x', t) + \sum_j \frac{\partial G_i'}{\partial \dot{x}_j}\dot{u}_j + \sum_j \frac{\partial G_i'}{\partial x_j}u_j + R_i = 0, \qquad i = r+1, \cdots, n.$$

λ' as a coefficient of $\dot{x}'$ is to be regarded as an α parameter. Consequently, $\partial G_i'/\partial(\lambda\dot{x}_j)$, $\partial G_i'/\partial\dot{x}_j$ and $\partial G_i'/\partial x_j$ are now functions of t independent of λ and u. R_i contains the higher powers of $\lambda\dot{u}$, $\dot{u}$ and u.

The time interval $0 \leqq t \leqq t^*$ is now to be subdivided into intervals $0 = t_0 < t_1 < \cdots < t_m = t^*$ and in each of these subintervals a point t_k', $t_{k-1} \leqq t_k' \leqq t_k$ is chosen at which we evaluate the coefficients

$$\partial G_i'/\partial(\lambda\dot{x}_j) = A_{ij}, \qquad i, j = 1, \cdots, r \tag{2.12}$$

$$\partial G_i'/\partial\dot{x}_j = B_{ij}, \qquad i, j = 1, \cdots, n \tag{2.12'}$$

$$\partial G_i'/\partial x_j = C_{ij}, \qquad i, j = 1, \cdots, n. \tag{2.12''}$$

These permit us to rewrite Eqs. (2.11) in the form

$$0 = G_i'(\lambda\dot{x}', \dot{x}', x', t) + \sum_j A_{ij}\lambda\dot{u}_j + \sum_j B_{ij}\dot{u}_j + \sum_j C_{ij}u_j + \eta \cdot\left[\sum_j\left(\frac{\partial G_i'}{\partial(\lambda\dot{x}_j)} - A_{ij}\right)\lambda\dot{u}_j + \sum_j\left(\frac{\partial G_i'}{\partial\dot{x}_j} - B_{ij}\right)\dot{u}_j + \sum_j\left(\frac{\partial G_i'}{\partial x_j} - C_{ij}\right)u_j + R_i\right], \qquad i = 1, \cdots, r \tag{2.13}$$

$$0 = G_i'(\dot{x}', x', t) + \sum_j B_{ij}\dot{u}_j + \sum_j C_{ij}u_j + \eta \cdot\left[\sum_j\left(\frac{\partial G_i'}{\partial\dot{x}_j} - B_{ij}\right)\dot{u}_j + \sum_j\left(\frac{\partial G_i'}{\partial x_j} - C_{ij}\right)u_j + R_i\right], \qquad i = r+1, \cdots, n$$

where η is an α-like parameter which we introduce. For $\eta = 1$ Eq. (2.13) is equivalent to the original system, Eqs. (2.7), and for $\eta = 0$, the system of Eq. (2.13) is a system with piecewise constant coefficients.

LEMMA 1. *If Hypothesis* 2.1 *above is satisfied for the given problem originally, it is also satisfied by the system of Eq.* (2.13) *at* $\eta = 0$.

The conditions of Hypothesis 2.1 are to be interpreted as the statement that certain Jacobians are not zero and certain additional statements. If we set $\eta = 0$, the corresponding Jacobians are readily seen to coincide with certain values of the former set of Jacobians and hence are not zero. Since at $\eta = 0$ we have a linear system in the u's, the existence of solutions at specific points is readily established. This establishes the Lemma.

The fact that Hypothesis 2.1 is satisfied for $\eta = 0$ and the fact that Eq. (2.13) for $\eta = 0$ is a linear system insures that given any initial condition $\dot{u}_{1,0}, \cdots, \dot{u}_{r,0}, u_{1,0}, \cdots, u_{n,0}$ we can find a solution corresponding to $\eta = 0$.

Furthermore, since under these circumstances the hypotheses of Theorem 1 above apply to this system and this solution, the $\dot{u}_1, \cdots, \dot{u}_r$ and $u_1, \cdots, u_n$ are analytic functions of η, α and β which, for a given value of $t = t^*$, are analytic for a neighborhood around zero. If this neighborhood contains a region for which $\eta = 1$ we shall say the set of initial conditions $\dot{u}_{1,0}, \cdots, \dot{u}_{r,0}, u_{1,0}, \cdots, u_{n,0}$ is in case 1.

We suppose now we are dealing with a set of initial conditions in case 1.

η and λ' are to be regarded as additional α parameters and we know by Theorem 1, that the $u_1, \cdots, u_n$ are analytic in all α and β for any given value of t in the specified interval. Thus $u_i(t, \lambda, \alpha, \beta)$ can be expanded in a power series around $\alpha, \beta = 0$ for any fixed value of λ, i.e.

$$u_i(t, \lambda, \alpha, \beta) = u_i(t, \lambda, 0, 0) + \sum u_{i,\alpha}\alpha + \sum u_{i,\beta}\beta + \cdots \tag{2.14}$$

where the coefficients, say, w_i are the partials of u_i relative to some combination of α's and β's. The w_i's, $i = 1, \cdots, n$ for a fixed combination of α's and β's, say, of the kth order satisfy a system of differential equations

$$\begin{aligned} 0 &= \textstyle\sum_j A_{ij}\lambda\dot{w}_j + \sum_j B_{ij}\dot{w}_j + \sum_j C_{ij}w_j + S_i, \qquad i = 1, \cdots, r \\ 0 &= \textstyle\sum_j B_{ij}\dot{w}_j + \sum_j C_{ij}w_j + S_i, \qquad i = r+1, \cdots, n \end{aligned} \tag{2.15}$$

where the S_i's involve partial derivatives of the u_i's of order lower than k. This, of course, is a system which can be solved successively for the various values of k. To do so we must first solve the homogeneous system

$$\begin{aligned} 0 &= \textstyle\sum_j A_{ij}\lambda\dot{w}_j + \sum_j B_{ij}\dot{w}_j + \sum_j C_{ij}w_j, \qquad i = 1, \cdots, r \\ 0 &= \textstyle\sum_j B_{ij}\dot{w}_j + \sum_j C_{ij}w_j, \qquad i = r+1, \cdots, n. \end{aligned} \tag{2.16}$$

This system is not one with constant coefficients but one in which the coefficients are constant over successive time intervals. Nevertheless we can construct $n + r$ linearly independent solutions and use these in turn to solve the nonhomogeneous systems of Eq. (2.15).

2.4. The Sensitivity Equations. We first consider the system of Eq. (2.16) for a fixed interval. Since this is a system with constant coefficients, we consider

the indicial equation which can be written in determinantal form as

$$| A_{ij}\lambda m^2 + B_{ij}m + C_{ij} | = 0. \tag{2.17}$$

(We let $A_{ij} = 0$ for $i = r + 1, \cdots, n$.)

If we set $\lambda = 0$ in the above, we obtain an nth order polynomial

$$| B_{ij}m + C_{ij} | = 0. \tag{2.18}$$

The matrix $B = \| B_{ij} \|$ corresponds to the Jacobian of the original system of Eq. (2.1) although it is not precisely equal to it as long as λ' is not zero. In view of this though, it is reasonable to assume that $\| B_{ij} \|$ is not singular. If we multiply the determinant of Eq. (2.18) by the determinant of the inverse matrix B^{-1} we have a characteristic equation

$$| mI + B^{-1}C | = 0$$

which clearly is of the nth degree in m. ($C = \| C_{ij} \|$.)

Assumption 1. $B = \| B_{ij} \|$ *is not singular.*

Assumption 2. *Equation* (2.18) *does not have multiple roots.*

LEMMA 2. *Assumption* 1 *implies that Eq.* (2.18) *is an equation of the nth degree.*

LEMMA 3. *There is a* λ *neighborhood of* $\lambda = 0$ *such that corresponding to each root* m_0 *of Eq.* (2.18) *which is not a multiple root of Eq.* (2.18) *there exists a root* $\mu(\lambda)$ *of Eq.* (2.17) *such that* $\mu(0) = m_0$ *and* $\mu(\lambda)$ *is analytic in* λ *at* $\lambda = 0$.

Proof. Consider

$$L(\lambda, m) = | A_{ij}\lambda m^2 + B_{ij}m + C_{ij} | \tag{2.19}$$

and consider Eq. (2.17), i.e., $L = 0$ as an equation intended to determine m as a function of λ. If we set $\lambda = 0$, m_0 is a value of m satisfying $L(0, m_0) = 0$ and we may apply the usual implicit function theorem to this situation and obtain the desired result provided $\partial L/\partial m \neq 0$ at $\lambda = 0$, $m = m_0$. But L reduces to Eq. (2.18) for $\lambda = 0$ and the statement $\partial L/\partial m \neq 0$ is immediately equivalent to the statement that m_0 is not a double root of Eq. (2.18).

Q. E. D.

(Multiple roots m_0 lead to expansions in fractional powers of λ rather than a simple power series. Otherwise the results are analogous but it should be appreciated that small variations in λ yield far larger variations in (say) the square root of λ. Consequently, in the multiple root case, the root $\mu(\lambda)$ depends much more sensitively on λ than in the analytic case for small values of λ.)

Under Assumptions 1 and 2 then we have n roots $\mu_1(\lambda), \cdots, \mu_n(\lambda)$ of Eq. (2.17) associated with corresponding roots of Eq. (2.18). [One might well point out here that in the case of no λ errors we could still apply the process of the previous section in order to obtain a system of equations with piecewise constant coefficients. The result would be Eq. (2.18). For these we would have the n roots $m_1, \cdots, m_n$ and correspondingly n solutions. The introduction of the λ errors has affected these only slightly under Assumptions 1 and 2. We have n

analytic functions of λ, $\mu_1(\lambda), \cdots, \mu_n(\lambda)$ each of which reduces to an m at $\lambda = 0$. The same holds for the corresponding solutions.]

Thus these n solutions correspond to relatively minor variations of the basic situation introduced by the λ error.

However, there are other solutions which must be investigated.

LEMMA 4. *Equation* (2.17) *is of degree* $n + r$ *in* m.

Proof. Consider first the determinant of the first r rows and r columns of the determinant in Eq. (2.17).

$$| A_{ij}\lambda m^2 + B_{ij}m + C_{ij} | \qquad i, j = 1, \cdots, r \tag{2.20}$$

This determinant is a polynomial in m of degree at most $2r$. Since the coefficient of m^{2r} is precisely the determinant $| A_{ij} | \lambda^{2r}$ and the last is not zero by Hypothesis 1, we see that this polynomial is of the $2r$ degree. Now one can also readily show that any other rth row determinant from the first r rows of the determinant in Eq. (2.17) is of degree $\leqq 2r - 1$ since only the first r columns are of the second degree in m.

In the last $n - r$ rows we take the last $n - r$ columns and consider the corresponding minor

$$| B_{ij}m + C_{ij} |, \qquad i, j = n - r + 1, \cdots, n. \tag{2.21}$$

This is of degree $\leqq n - r$ with $| B_{ij} |$ as the coefficient of m^{n-r}. Hypothesis 1 states that this determinant is not zero and consequently Eq. (2.21) is of degree $n - r$. Any other minor from the last $n - r$ rows is of degree $\leqq n - r$.

Apply then the Laplace expansion of L for the first r rows. This expansion contains the product of the Eqs. (2.20) and (2.21) which is of the $n + r$ degree. Since the degree of the other minors of the first r rows is less than $2r$ and the degree of the minors of the last $n - r$ rows does not exceed $n - r$, the degree of any other product in this expansion is less than $n + r$. Hence L is of degree $n + r$.

Thus there are r solutions of Eq. (2.17) which we must still obtain. Multiply Eq. (2.17) by λ^n. This will permit us to multiply each element in the determinant by λ,

$$| A_{ij}(\lambda m)^2 + B_{ij}(\lambda m) + C_{ij}\lambda | = 0. \tag{2.22}$$

Let $\nu = \lambda m$. Then Eq. (2.22) becomes

$$| A_{ij}\nu^2 + B_{ij}\nu + C_{ij}\lambda | = 0. \tag{2.23}$$

Equation (2.23) is, of course, of the same degree in ν as Eq. (2.17) was in m, i.e., of degree $n + r$. Let us now consider Eq. (2.23) for $\lambda = 0$, obtaining

$$| A_{ij}\nu^2 + B_{ij}\nu | = 0. \tag{2.24}$$

We can factor a ν from each column of this determinant which can then be written as

$$\nu^n | A_{ij}\nu + B_{ij} | = 0. \tag{2.25}$$

Now

$$|A_{ij}\nu + B_{ij}| = 0 \tag{2.26}$$

is an equation of degree r in ν. Suppose $\nu_{1,0}, \cdots, \nu_{r,0}$ are the r roots of Eq. (2.26).

Assumption 3. *Equation* (2.26) *does not have multiple roots.*

Assumption 1 implies that no $\nu_{k,0}$ is zero.

LEMMA 5. *Under Assumptions* 1 *and* 3, *there exists a* λ *neighborhood of* $\lambda = 0$ *such that for each root* $\nu_{k,0}$ *of Eq.* (2.26), *there exists an analytic solution* $\nu(\lambda)$ *of Eq.* (2.23).

Proof. The proof of this is similar to that of Lemma 3. If we retrace the steps from Eq. (2.23) to Eq. (2.26) in reverse order we find that each of the solutions of Eq. (2.26) yields a solution of Eq. (2.23) for $\lambda = 0$. Assumption 1 shows that $\nu = 0$ is not one of the $\nu_{k,0}$ and by Assumption 3, the $\nu_{k,0}$ are distinct. Consequently, no $\nu_{k,0}$ is a multiple root of Eq. (2.23) at $\lambda = 0$, and $\partial L/\partial\nu \neq 0$ at each of the values $\lambda = 0$, $\nu = \nu_{k,0}$. The implicit function theorem can now be applied to obtain the result specified.

Corresponding to the r solutions $\nu(\lambda)$ of Eq. (2.23) there are r solutions $m = \nu(\lambda)/\lambda$ of Eq. (2.22) and consequently r solutions of Eq. (2.17).

The situation for any multiple root of Eq. (2.26) is analogous to that for any multiple root of Eq. (2.21). We summarize our results in the following theorem.

THEOREM 2. *Under Hypothesis* 2.1 *above and Assumptions* 1, 2, *and* 3 *above, for each interval of constancy the indicial equation Eq.* (2.17) *of the system Eq.* (2.16) *possesses n roots* $\mu_1(\lambda), \cdots, \mu_n(\lambda)$ *analytic in* λ *and r roots* $\nu_1(\lambda)/\lambda, \cdots, \nu_r(\lambda)/\lambda$ *where* $\nu_1(\lambda), \cdots, \nu_r(\lambda)$ *are analytic in* λ. *If Assumptions* 1 *or* 2 *do not hold, the* μ*'s or the* ν*'s can be expressed in fractional powers of* λ.

2.5. The Solution for a Single Interval. To each solution $\mu_j(\lambda)$ of the theorem of the previous section, we can find a solution of Eq. (2.16) in the form

$$u_i^{\,j} = D_{ij}e^{\mu_j(\lambda)t} \tag{2.27}$$

on the interval of constancy. The D_{ij} are the non-trivial solutions of the homogeneous system

$$\sum_k (A_{ik}\lambda\mu_j^{\,2} + B_{ik}\mu_j + C_{ik})D_{kj} = 0, \qquad i = 1, \cdots, n \tag{2.28}$$

whose determinant is zero. ($A_{ij} = 0$, $i = r + 1, \cdots, n$.) Similarly for each of the $\nu_j(\lambda)$ we have solutions in the form

$$v_i^{\,j} = E_{ij}e^{(\nu_j(\lambda)/\lambda)t} \tag{2.29}$$

where for each j

$$\sum_k (A_{ik}\nu_j^{\,2} + B_{ik}\nu_j + \lambda C_{ik})E_{kj} = 0. \tag{2.30}$$

Notice that the Eq. (2.28) and (2.30) which determine D_{kj} and E_{kj} when a

suitable normalization is supposed, are well defined at $\lambda = 0$ and are analytic in λ at $\lambda = 0$.

The general solution of the system

$$\sum_k A_{ik}\lambda\dot{u}_k + \sum_k B_{ik}\dot{u}_k + \sum_k C_{ik}u_k = 0 \tag{2.31}$$

is a linear combination with constant coefficients of Eqs. (2.27) and (2.29). In vector notation

$$u = \sum_{j=1}^{n} U_j u^j + \sum_{k=1}^{r} V_k v^k. \tag{2.32}$$

Here u^j and v^k stand for vectors with elements given by Eqs. (2.27) and (2.29) above. Now let U stand for the vector with elements U_j, V for that with elements V_j, M for the diagonal matrix with elements $\mu_j(\lambda)\delta_{ij}$, N the rth order diagonal matrix with elements $\nu_j(\lambda)\delta_{ij}$. Then Eq. (2.32) may be written

$$u = D(\exp Mt)U + E[\exp((N/\lambda)t)]V. \tag{2.33}$$

(D is the matrix $\| D_{ij} \|$, E is an $n \times r$ matrix $\| E_{ij} \|$.)

Let D^r and E^r denote the matrix of the first r rows of D and E respectively. Then the initial conditions for the solution u can be written in vector form as

$$u^0 = DU + EV, \qquad \lambda\dot{u}^0 = \lambda D^r MU + E^r NV. \tag{2.34}$$

U and V as vectors can also be considered as one column matrices. Similarly, u^0 and $\dot{u}^0$ are one column matrices. Thus,

$$U = D^{-1}u^0 - D^{-1}EV$$

and

$$\lambda(\dot{u}^0 - D^r MD^{-1}u^0) = (E^r N - \lambda D^r MD^{-1}E)V.$$

Now let

$$T = E^r N - \lambda D^r MD^{-1}E. \tag{2.35}$$

Then

$$V = \lambda T^{-1}(\dot{u}^0 - D^r MD^{-1}u^0) \tag{2.36}$$

$$U = D^{-1}u^0 - \lambda D^{-1}ET^{-1}(\dot{u}^0 - D^r MD^{-1}u^0). \tag{2.37}$$

Notice V is zero and $U = D^{-1}u^0$ provided

$$\dot{u}^0 = D^r MD^{-1}u^0. \tag{2.38}$$

LEMMA 6. *Equation* (2.38) *is the necessary and sufficient condition that Eq.* (2.33) *depend analytically on* λ *at* $\lambda = 0$.

These, of course, all apply to an interval of t on which the coefficients A_{ij}, B_{ij}, C_{ij} are constant. Notice, however, that if $\Re(\nu_j) < 0$ (real part of ν_j negative) and of finite size while λ is small the non-analytic terms of Eq. (2.33) are negligible for non-zero positive values of t of finite size.

Equations (2.33), (2.36), and (2.37) can be combined to yield

$$u = [D\exp(Mt)D^{-1}]u^0 + \lambda[E\exp((N/\lambda)t) - D\exp(Mt)D^{-1}E]T^{-1} \cdot (\dot{u}^0 - D^rMD^{-1}u^0) \quad (2.39)$$

$$= \exp(\bar{M}t)u^0 + \lambda[E\exp((N/\lambda)t) - \exp(\bar{M}t)E]T^{-1}(\dot{u}^0 - D^rD^{-1}\bar{M}u^0)$$

where $\bar{M} = DMD^{-1}$.

2.6. The Continuation of the Solution. Equation (2.39) applies to each interval of constancy and relates a solution of the homogeneous system Eq. (2.16), u, to its initial values u^0 and $\dot{u}^0$ for the interval. u^0 is an n-dimensional vector, $\dot{u}^0$ is r-dimensional. A subscript following a colon will indicate the interval, i.e. $u_{j:k}$ is the jth component of the solution vector on the kth interval. Equation (2.39) can then be written

$$u_{:k} = \exp[\bar{M}_{:k}(t - t_{k-1})]u^0_{:k} + \lambda\{E_{:k}\exp[(N_{:k}/\lambda)(t - t_{k-1})] - \exp[\bar{M}_{:k}(t - t_{k-1})E_{:k}\}\ T^{-1}_{:k}(\dot{u}^0_{:k} - D^r_{:k}D^{-1}_{:k}\bar{M}_{:k}u^0_{:k}) \quad (2.40)$$

which implies

$$\dot{u}_{:k} = \bar{M}_{:k}\exp[\bar{M}_{:k}(t - t_{k-1})]u^0_{:k} + \{E_{:k}N^r_{:k}\exp[(N_{:k}/\lambda)(t - t_{k-1})] - \lambda\bar{M}_{:k}\exp[\bar{M}_{:k}(t - t_{k-1})]E_{:k}\}\ T^{-1}_{:k}(\dot{u}^0_{:k} - D^r_{:k}D^{-1}_{:k}\bar{M}_{:k}u^0_{:k}). \quad (2.41)$$

Let H denote the $r \times n$ dimensional matrix whose first $r \times r$ minor matrix is the identity and the rest zero. Then

$$D^r = HD \text{ and } D^rD^{-1} = H.$$

It is now desirable to obtain the relation between $u^0_{:k-1}$ and $\dot{u}^0_{:k-1} - H\bar{M}_{:k-1}u^0_{:k-1}$ and $u^0_{:k}$ and $\dot{u}^0_{:k} - H\bar{M}_{:k}u^0_{:k}$, (cf. Eq. (2.38)). Let $\Delta t_{k-1} = t_{k-1} - t_{k-2}$. The given solution is continuous at $t = t_{k-1}$ if

$$u^0_{:k} = \exp(\bar{M}_{:k-1}\Delta t_{k-1})u^0_{:k-1} + \lambda\{E_{:k-1}\exp[(N_{:k-1}/\lambda)\Delta t_{k-1}] - \exp(\bar{M}_{:k-1}\Delta t_{k-1})E_{:k-1}\}\ T^{-1}_{:k-1}(\dot{u}^0_{:k-1} - H\bar{M}_{:k-1}u^0_{:k-1}). \quad (2.42)$$

The derivatives of the first r of the u_j's are continuous if

$$\dot{u}^0_{:k} = H\bar{M}_{:k-1}\exp(\bar{M}_{:k-1}\Delta t_{k-1})u^0_{:k-1} + H\{E_{:k-1}N_{:k-1}\exp[(N_{:k-1}/\lambda)\Delta t_{k-1}] - \lambda\bar{M}_{:k-1}\exp(\bar{M}_{:k-1}\Delta t_{k-1})E_{:k-1}\} \cdot T^{-1}_{:k-1}(\dot{u}^0_{:k-1} - H\bar{M}_{:k-1}u^0_{:k-1}). \quad (2.43)$$

Now let $v^0_{:k} = \dot{u}^0_{:k} - H\bar{M}_{:k}u^0_{:k}$. Equation (2.42) can then be written as

$$u^0_{:k} = \exp(\bar{M}_{:k-1}\Delta t_{k-1})u^0_{:k-1} + \lambda\{E_{:k-1}\exp[(N_{:k-1}/\lambda)\Delta t_{k-1}] - \exp(\bar{M}_{:k-1}\Delta t_{k-1})E_{:k-1}\}T^{-1}_{:k-1}v^0_{:k-1} \quad (2.44)$$

and

$$v^0{}_{:k} = H\{(\bar{M}_{:k-1} - \bar{M}_{:k}) \exp(\bar{M}_{:k-1}\Delta t_{k-1})u^0{}_{:k-1} + [E_{:k-1}(N_{:k-1} - \lambda\bar{M}_{:k}) \exp((N_{:k-1}/\lambda)\Delta t_{k-1}) - \lambda(\bar{M}_{:k-1} - \bar{M}_{:k}) \exp(\bar{M}_{:k-1}\Delta t_{k-1})E_{:k-1}]T^{-1}{}_{:k-1}v^0{}_{:k-1}\}. \tag{2.45}$$

Thus, Eqs. (2.44) and (2.45) permit us to continue a solution of the homogeneous solution, through intervals. For suitably stable matrices, the exponential matrices should cause such a solution to decay.

In general, the solution Eq. (2.40) consists of two parts. One part is analytic in λ, i.e., the term involving $\exp \bar{M}(t - t_{k-1})$. Normally one would expect this to decay slowly. On the other hand, the matrix $\exp (N(\lambda)/\lambda)(t - t_{k-1})$, if it is properly stable, should have very large negative real parts for the characteristic roots and hence should decay quickly. This is, however, non-analytic in λ. The possibility that $\exp \bar{M}(t - t_{k-1})$ is unstable must be considered. (Cf. Sec. 1.5.)

2.7 Construction of the Solution. The results of the previous section can be used to solve the system of Eq. (2.15) as follows: the solution desired is that for the non-homogeneous system, Eq. (2.15), which at $t = t_0$ has zero values for $u_1, \cdots, u_n, \dot{u}_1, \cdots, \dot{u}_r$. For the first interval $t_0 \leqq t \leqq t_1$ a solution $y_{:1}$ will be found of Eq. (2.15). This will have certain values at t_1 which can be used as the initial values for a solution of the homogeneous equation for the second interval and this solution can be continued as a solution of the homogeneous solution for the remaining intervals. In general, let $y_{:k}$ denote the vector solution $y_{1:k}, \cdots, y_{n:k}$ which is zero for the intervals preceding the kth, is a solution of Eq. (2.15) on the kth interval with initial values zero and is continued continuously by the method of the previous section as a solution of the homogeneous equation over the remaining intervals.

Since it is obvious that the desired solution $u_1, \cdots, u_n$ is the sum of the $y_{1:k}, \cdots, y_{n:k}$ for all intervals I_k, it remains only to solve the problem of finding, for a given interval I_k, the solution of the non-homogeneous system Eq. (2.15) which at the lower endpoint has zero initial values. Thus again one needs to consider only a single interval and it is not necessary to make explicit reference to the interval.

It is desirable to solve the first r equations of Eq. (2.15) for $\lambda\dot{z}_j$, the remaining for $\dot{z}_j$. Thus,

$$\begin{aligned} \lambda\dot{z}_i &= \textstyle\sum_{j=1}^{n} \beta'_{ij}\dot{z}_j + \sum_{j=1}^{n} \gamma'_{ij}z_j + s'_i, \qquad i = 1, \cdots, r \\ \dot{z}_i &= \textstyle\sum_{j=1}^{n} \gamma'_{ij}z_j + s'_i, \qquad i = r+1, \cdots, n \end{aligned} \tag{2.46}$$

where s'_i is a linear combination of the original non-homogeneous terms. It is desirable to use the last equations to eliminate $\dot{z}_{r+1}, \cdots, \dot{z}_n$ from the first r which will then become

$$\begin{aligned} \lambda\dot{z}_i &= \textstyle\sum_{j=1}^{r} \beta_{ij}\dot{z}_j + \sum_{j=1}^{n} \gamma_{ij}z_j + s_i, \qquad i = 1, \cdots, r \\ \dot{z} &= \textstyle\sum_{j=1}^{n} \gamma_{ij}z_j + s_i, \qquad i = r+1, \cdots, n. \end{aligned} \tag{2.47}$$

One may recall that the customary method of finding a solution of the non-homogeneous equation Eq. (2.47), is to consider the expression Eq. (2.32) with the constants U_j and V_k replaced by functions $\varphi_j(t)$ and $\psi_k(t)$ of t

$$\begin{aligned} z_i &= \textstyle\sum_{j=1}^{n} \varphi_j(t) u_i^{\,j} + \sum_{k=1}^{r} \psi_k(t) v_i^{\,k} \\ &= \textstyle\sum_{j=1}^{n} D_{ij} \varphi_j(t) e^{\mu_j t} + \sum_{k=1}^{r} E_{ij} \psi_j(t) e^{(\nu_j/\lambda)t}, \end{aligned} \tag{2.48}$$

where $u_i^{\,j}$ and $v_i^{\,j}$ are given by Eqs. (2.27) and (2.29). Since Eq. (2.47) is of the second order in z_i, $i = 1, \cdots, r$, one must introduce auxiliary conditions

$$\textstyle\sum_{j=1}^{n} \dot{\varphi}_j u_i^{\,j} + \sum_{k=1}^{r} \dot{\psi}_k v_i^{\,k} = 0, \qquad i = 1, \cdots, r. \tag{2.49}$$

Since the $u_i^{\,j}$ and $v_i^{\,k}$ are solutions of the homogeneous system, one can readily show that Eq. (2.47) is equivalent to

$$\begin{aligned} \lambda\left[\textstyle\sum_{k=1}^{r} \dot{\psi}_k v_i^{\,k} + \sum_{j=1}^{n} \dot{\varphi}_j \dot{u}_i^{\,j}\right] &= s_i, \qquad i = 1, \cdots, r \\ \textstyle\sum_{k=1}^{r} \dot{\psi}_k v_i^{\,k} + \sum_{j=1}^{n} \dot{\varphi}_j u_i^{\,j} &= s_i, \qquad i = r+1, \cdots, n. \end{aligned} \tag{2.50}$$

Combining these with Eqs. (2.27) and (2.29), one obtains

$$\begin{aligned} \textstyle\sum_{k=1}^{r} E_{ik} \nu_k \dot{\psi}_k e^{(\nu_k/\lambda)t} + \sum_{j=1}^{n} D_{ij} \lambda \mu_j \dot{\varphi}_j e^{\mu_j t} &= s_i, \qquad i = 1, \cdots, r \\ \textstyle\sum_{k=1}^{r} E_{ik} \dot{\psi}_k e^{(\nu_k/\lambda)t} + \sum_{j=1}^{n} D_{ij} \dot{\varphi}_j e^{\mu_j t} &= 0, \qquad i = 1, \cdots, r \\ \textstyle\sum_{k=1}^{r} E_{ik} \dot{\psi}_k e^{(\nu_k/\lambda)t} + \sum_{j=1}^{n} D_{ij} \dot{\varphi}_j e^{\mu_j t} &= s_i, \qquad i = r+1, \cdots, n. \end{aligned} \tag{2.51}$$

Now, if one introduces $\Psi_k = \dot{\psi}_k e^{(\nu_k/\lambda)t}$, $\Phi_j = \dot{\varphi}_j e^{\mu_j t}$ Eq. (2.51) becomes

$$\begin{aligned} \textstyle\sum_{k=1}^{r} E_{ik} \nu_k \Psi_k + \sum_{j=1}^{n} D_{ij} \lambda \mu_j \Phi_j &= s_i, \qquad i = 1, \cdots, r \\ \textstyle\sum_{k=1}^{r} E_{ik} \Psi_k + \sum_{j=1}^{n} D_{ij} \Phi_j &= 0, \qquad i = 1, \cdots, r \\ \textstyle\sum_{k=1}^{r} E_{ik} \Psi_k + \sum_{j=1}^{n} D_{ij} \Phi_j &= s_i, \qquad i = r+1, \cdots, n. \end{aligned} \tag{2.52}$$

This is a linear system of equations with constant coefficients on the Φ_k and Ψ_k as far as t is concerned. Furthermore, for $\lambda = 0$, the determinant is in the form of the product of the determinant of E' and D evaluated at $\lambda = 0$ (cf. above) and hence is not zero. Since this determinant is a continuous function of λ, one may suppose the existence of a range of λ around $\lambda = 0$, where this determinant is not zero.

Solving for Ψ_k and Φ_k expresses these also as linear combinations with constant coefficients of the original inhomogeneous terms.

$$\Psi_k = s_k^{*}(t), \qquad \Phi_j = s_j''(t). \tag{2.53}$$

The coefficients depend analytically on λ and thus,

$$\dot{\psi}_k = e^{-(\nu_k/\lambda)t} s_k^{*}(t), \qquad \dot{\varphi}_j = e^{-\mu_j t} s_j''(t). \tag{2.54}$$

Since at the initial point t_{k-1}, ψ_j and φ_j are zero one has for the desired solution of Eq. (2.46)

$$z_i(t) = \sum_{j=1}^{n} D_{ij} e^{\mu_j t} \int_{t_{k-1}}^{t} e^{-\mu_j \tau} s_j''(\tau)\, d\tau + \sum_{j=1}^{r} E_{ij} e^{(\nu_j/\lambda)t} \int_{t_{k-1}}^{t} e^{-(\nu_j/\lambda)\tau} s_j^{*}(\tau)\, d\tau \tag{2.55}$$

valid in the interval I_k.

We conclude therefore:

LEMMA 7. *The solution $y_{i,k}$ defined at the beginning of this section has the form of Eq.* (2.55) *on the* k*th interval. The quantities and functions* D_{ij}, μ_j, $s_j^{\prime\prime}(\tau)$, E_{ij}, ν_j, $s^*(\tau)$ *depend analytically on* λ *in some neighborhood of* $\lambda = 0$.

For each interval then the solution $u_1, \cdots, u_n$ of Eq. (2.15) is in the form $\sum_{j=1}^{k} y_{ij}$. For $j < k$, the y_{ij} are linear combinations of terms in the form $e^{\mu_i t}$ and $e^{(\nu_j/\lambda)t}$. Normally one would expect that the terms $e^{\mu_i t}$ would decay slowly and those in the form $e^{(\nu_j/\lambda)t}$ would decay quickly. The above lemma shows that $y_{i,k}$ may also be divided into two terms similar in nature to these two types. Thus it is clear that while the λ errors do introduce errors in the solution, non-analytic in λ, these errors decay quickly if the ν_j have negative real parts.

2.8. Concluding Discussion. It is now possible to set up a general theory of the error, including all types of errors, α, β and λ. As in the above we "linearize" the problem by introducing the extra α parameters, λ' and η. Now for λ fixed we can expand the error function $u(t, \lambda, \alpha, \beta)$ in terms of α and β. Thus

$$u(t, \lambda, \alpha, \beta) = u(t, \lambda, 0, 0) + \sum_{\gamma=\alpha,\beta} u_\gamma(t, \lambda, 0, 0)\gamma + \cdots . \tag{2.56}$$

The above argument yields the coefficients of this expansion $w(t, \lambda, 0, 0)$ as solutions of Eq. (2.15).

The discussion of the previous sections show that in each case there is a favorable situation where the expression for the partial

$$\frac{\partial^r u}{\partial_{\gamma_1}^{r_1} \cdots \partial_{\gamma_u}^{r_u}} \tag{2.57}$$

can be expressed as a sum of two terms, one of which is analytic in λ, while the other term, based on $\nu_j(\lambda)$ with negative real parts, disappears rapidly. Consequently, we have a similar resolution for $u(t, \lambda, \alpha, \beta)$ in the favorable case in which the $\nu_j(\lambda)$ have negative real parts. If the real part of $\nu_j(\lambda)$ is positive, the $\nu_j(\lambda)$ terms for λ small will be dominant and very large and normally one would expect the machine solution to be useless.

The above discussion is based on a number of assumptions. One of these is that relative to the parameter η, the solution is in case 1. Another is that the order of the system increases by no more than 1, when introduced into the machine. However, this last assumption was introduced merely for convenience. One can readily indicate a procedure which is applicable in the higher order case. The essential part of this discussion is the generalization of the situation represented by Hypothesis 2.1 above. The given system $F_i = 0$ is supposed to be in the first order case. Let s denote the order of the highest derivative which appears in the machine equations $G_i = 0$. If our given system $G_i = 0$ implies $n - r_s$ relations between the derivatives of lower order, we suppose that we can separate out r_s equations involving the s-order derivatives and that in these in turn we can eliminate all but r_s s-order derivatives. In the remaining $n - r_s$ relations, we suppose that we have $n - r_s - r_{s-1}$ relations of order less than $s - 1$. A similar procedure then permits us to separate out r_{s-1} relations on r_{s-1} derivatives of order $s - 1$.

Thus we can obtain, after a suitable re-enumeration of the dependent variables, a system equivalent to the original broken up into s sets of equations. The first set consists of r_s equations on the s-order derivatives of $x_1, \cdots, x_{r_s}$ and do not involve any further derivatives of order s. The next set consists of r_{s-1} equations on the $s-1$ derivatives of $x_{r_s+1}, \cdots, x_{r_s+r_{s-1}}$ and contains no higher derivatives and no other derivatives of order $s-1$. A similar situation holds for the remaining sets of equations. The parameter λ is most effectively introduced by writing the functions as depending on $\lambda^{s-1}x_j^{(s)}$, $\lambda^{s-2}x_j^{(s-1)}$, etc.

The linearization process of §2.3 of this chapter may now be applied. Setting $\lambda = 0$ will again yield n values $\mu_{1,0}, \cdots, \mu_{n,0}$ which determine functions $\mu_j(\lambda)$ analytic in λ at $\lambda = 0$ (cf. §2.4 above). On the other hand, one can also show that one has a total of $r_2 + 2r_3 + \cdots + (s-1)r_s$ extra ν roots of the equation equivalent to Eq. (2.23) of §2.4 above. From this point on, the results in the higher order case are precisely similar to those in the case in which the order is raised one.

Returning now to the case in which the order is raised one, we wish to discuss the alternative to the assumption above that our problem is in case 1. Normally one would want the R_i of Eq. (2.11) of §2.3 above to be small and this requires that the $\dot{u}_1, \cdots, \dot{u}_r$ be initially small since it is reasonable to expect that the $u_1, \cdots, u_n$ are small. However, unless special precautions are taken, the $\dot{u}_1, \cdots, \dot{u}_r$ will not in general be small and their size can lead to convergence difficulties or even to slow convergence in case 1.

There is a process which permits one to study the situation which arises in the case of large $\dot{u}_{1,0}, \cdots, \dot{u}_{r,0}$, i.e., the case where initially the machine solution and the true solution have different rates of change.

Consider the equations

$$G_i(\lambda\dot{u}_1, \cdots, \lambda\dot{u}_r, \dot{u}_1, \cdots, \dot{u}_n, u_1, \cdots, u_n, t) = 0, \qquad i = 1, \cdots, r$$

$$G_i(\dot{u}_1, \cdots, \dot{u}_n, u_1, \cdots, u_n, t) = 0,$$

and let us change the independent variable in u from t to a variable $\tau = \lambda^{-1}t$ and let the dot refer to differentiation relative to τ. The result is

$$G_i(\lambda^{-1}\dot{u}_1, \cdots, \lambda^{-1}\dot{u}_r, \lambda^{-1}\dot{u}_1, \cdots, \lambda^{-1}\dot{u}_n, u_1, \cdots, u_n, \lambda\tau) = 0$$

$$G_i(\lambda^{-1}\dot{u}_1, \cdots, \lambda^{-1}\dot{u}_n, u_1, \cdots, u_n, \lambda\tau) = 0.$$

The effect of the λ^{-1} is to emphasize those variables which it multiplies and in general one can consider this system equivalent to

$$H_i(\dot{u}_1, \cdots, \dot{u}_r, \dot{u}_1, \cdots, \dot{u}_n, \lambda, u_1, \cdots, u_n, \lambda\tau) = 0, \qquad i = 1, \cdots, r$$

$$H_i(\dot{u}_1, \cdots, \dot{u}_n, \lambda, u_1, \cdots, u_n, \lambda\tau) = 0, \qquad i = r+1, \cdots, n,$$

in which λ plays the role of a parameter. Thus, if we set $\lambda = 0$, we obtain relations

$$H_i^0(\dot{u}_1, \cdots, \dot{u}_r, \dot{u}_1, \cdots, \dot{u}_n) = 0, \qquad i = 1, \cdots, r$$

$$H_i^0(\dot{u}_1, \cdots, \dot{u}_n) = 0, \qquad i = r+1, \cdots, n.$$

These equations can be considered as a system of equations for $\dot{u}_1, \cdots, \dot{u}_n$. Now the favorable phenomena which is desired is that the solutions $\dot{u}_1, \cdots, \dot{u}_n$ of the above system should approach zero asymptotically as $\tau \to \infty$ and that the solutions be near zero after not too large a τ interval. Now suppose that λ is quite small. For the original variable t, this means that $\dot{u}_i$ becomes small after a small t interval and from then on one has a situation corresponding to small $u_1, \cdots, u_n$ and small $\dot{u}_1, \cdots, \dot{u}_r$. This process of examining the result of passing to the variable τ is basic in any discussion of λ errors and may be effectively used in case 1 as well.

In the above discussion, we have utilized only one parameter λ. If we have more than one type of integrator in the system, λ will be associated with the type which has the greatest time delay and other time delays will be represented by $a\lambda$ where a is small.

2.9. The Necessity for Interval by Interval Analysis. Given the differential equation

$$\lambda\ddot{y} + a(x)\dot{y} + b(x)y = 0 \tag{2.58}$$

with λ a parameter and $a(x)$, $b(x)$ analytic in the neighborhood of $x = 0$, it was hoped that the general solution of Eq. (2.58) could be written in the form

$$g(x, \lambda) = A \cdot X(x, \lambda) + B \cdot Y(x, \lambda)$$

where X and Y are analytic in x for a suitable interval and $Y(x, \lambda)$ is analytic in λ. Essentially this amounts to the requirement that Eq. (2.58) have one non-trivial solution which is analytic in λ. This, unfortunately, is not the case in general, as is shown in the following example.

Example: Consider the differential equation

$$\lambda\ddot{y} + \dot{y} + y/(1 + x) = 0. \tag{2.59}$$

Then, if there exists a solution $Y(x, \lambda)$ analytic in λ,

$$Y(x, \lambda) = u_0(x) + \lambda u_1(x) + \lambda^2 u_2(x) + \cdots \tag{2.60}$$

which can be differentiated with respect to x, twice, term by term, i.e., such that

$$\begin{aligned} \dot{Y}(x, \lambda) &= \dot{u}_0(x) + \lambda\dot{u}_1(x) + \lambda^2\dot{u}_2(x) + \cdots \\ \ddot{Y}(x, \lambda) &= \ddot{u}_0(x) + \lambda\ddot{u}_1(x) + \lambda^2\ddot{u}_2(x) + \cdots \end{aligned} \tag{2.61}$$

and $Y(0, 0) = 1$, then the λ radius of convergence of Eq. (2.60) for $x > 0$ is zero. (This is also true for $x < 0$.)

Proof. Substituting Eq. (2.61) in Eq. (2.59) we obtain the following set of differential systems:

$$\dot{u}_0 + u_0/(1 + x) = 0, \qquad u_0(0) = 1 \tag{2.62}$$

$$\dot{u}_n + u_n/(1 + x) = -\ddot{u}_{n-1}, \qquad u_n(0) = 0, \qquad n = 1, 2, \cdots. \tag{2.63}$$

By direct calculation,

$$u_0(x) = \frac{1!}{1+x}, \qquad u_1(x) = \frac{2!}{(1+x)^2} - \frac{2}{1+x}$$

$$u_2(x) = \frac{3!}{(1+x)^3} - \frac{4}{(1+x)^2} - \frac{2}{1+x}.$$

We shall now prove

$$(a) \qquad u_n(x) = \frac{(n+1)!}{(1+x)^{n+1}} - \sum_{k=1}^{n} \frac{a_k{}^n}{(1+x)^k}$$

$$(b) \qquad \sum_{k=1}^{n} a_k{}^n = (n+1)!$$

$$(c) \qquad a_k{}^n > 0 \qquad k = 1, \cdots, n.$$

The proof is by complete induction on n. Assume (a), (b), (c). Then substituting in Eq. (2.63):

$$\dot{u}_{n+1} + \frac{1}{1+x} u_{n+1} = -\frac{(n+1)(n+2)!}{(1+x)^{n+3}} + \sum_{k=1}^{n} \frac{k(k+1)}{(1+x)^{k+2}} a_k{}^n.$$

Assume

$$u_{n+1}(x) = \frac{a}{(1+x)^{n+2}} - \sum_{k=1}^{n+1} \frac{a_k{}^{n+1}}{(1+x)^k}.$$

Then

$$\begin{aligned}
\dot{u}_{n+1} + \frac{1}{1+x} u_{n+1} &= -\frac{(n+2)a}{(1+x)^{n+3}} + \sum_{k=1}^{n+1} \frac{k a_k{}^{n+1}}{(1+x)^{k+1}} \\
&\quad + \frac{a}{(1+x)^{n+3}} - \sum_{k=1}^{n+1} \frac{a_k{}^{n+1}}{(1+x)^{k+1}} \\
&= \frac{-(n+1)a}{(1+x)^{n+3}} + \sum_{k=1}^{n+1} \frac{(k-1)a_k{}^{n+1}}{(1+x)^{k+1}} \\
&= \frac{-(n+1)(n+2)!}{(1+x)^{n+3}} + \sum_{k=1}^{n} \frac{k(k+1)}{(1+x)^{k+2}} a_k{}^n.
\end{aligned}$$

Equate coefficients of $1/(1+x)^k$:

$$-(n+1)a = -(n+1)(n+2)!$$

$$(k-1)a_k{}^{n+1} = (k-1)k a_{k-1}^n, \qquad k = 2, \cdots, n+1$$

$$a_k{}^{n+1} = k a_{k-1}^n, \qquad k = 2, 3, \cdots, n+1$$

$$a = (n+2)!$$

$$a_k{}^{n+1} > 0, \qquad k = 2, 3, \cdots, n+1.$$

Let $a_1{}^{n+1} = a - \sum_{k=2}^{n+1} a_k{}^{n+1}$, then if we show $a_1{}^{n+1} > 0$, conditions (a), (b), (c)

will be satisfied. Now,

$$a_k^{n+1} \leq {}^{(n+1)}a_{k-1}^n, \qquad k = 2, 3, \cdots, n+1$$

$$\sum_{k=2}^{n+1} a_k^{n+1} < (n+1)\sum_{k=2}^{n+1} a_{k-1}^n = (n+1)\left(\sum_{i=1}^n a_i^n\right) = (n+1)(n+1)!$$

$$a_1^{n+1} > a - (n+1)(n+1)! = (n+2)! - (n+1)(n+1)!$$

$$= (n+1)! > 0.$$

Now suppose $x > 0$. Then since a_k^n is positive

$$\begin{aligned} u_n(x) &= (1+x)^{-n-1}\left[(n+1)! - \sum_{k=1}^n (1+x)^{n+1-k} a_k^n\right] \\ &< (1+x)^{-n-1}\left[(n+1)! - (1+x)\sum_{k=1}^n a_k^n\right] \\ &= -x(1+x)^{-n-1}(n+1)!. \end{aligned}$$

Thus for $x > 0$, $u_n(x)$ is more negative than $-x(n+1)!/(1+x)^{n+1}$ and the series $\sum_{n=0}^{\infty} u_n(x)\lambda^n$ has zero radius of convergence in λ. (A similar argument holds for $x < 0$.)

This example indicates that one cannot hope to break down the solutions of the linear variational equation into (say) an n-dimensional set of solutions analytic in λ at $\lambda = 0$ and an r-dimensional set which is non-analytic. The set of analytic solutions may have dimension less than n, even zero.

Thus, in order to obtain the long range part of the error, which in the more desirable cases contains the greater part of the error, one must use some procedure in addition to solving the variational equations, for instance, the process given above of breaking the interval into smaller intervals on which the coefficients may be considered to be constant. In addition, the latter process does yield a clear picture of the phenomena and even in the case when λ errors do not occur it may correspond to a desirable numerical procedure.

REFERENCES

[1] Brock, P. and Murray, F. J., "Planning and error analysis for the numerical solution of a test system of differential equations on the IBM sequence calculator," Project Cyclone, Reeves Instrument Corp., New York, N. Y.

[2] Bruevič, N. G., "On the accuracy of the fundamental formula of the theory of errors of a mechanism," *Bull.* Acad. Sci. USSR, Cl. Sci. Tech. [*Izvestiya* Akad. Nauk SSSR] **1944**, 1944, 545–558. (Russian)

[3] Bush, V., "The differential analyzer. A new machine for solving differential equations," *Jn.* of the Franklin Inst., **212**, 1931, 447–488.

[4] Bythovskiĭ, M. L., "The accuracy of mechanisms controlled by differential equations," *Izvestiya* Akad. Nauk SSSR. Otd. Tehn. Nauk **1947**, 1947, 1455–1512. (Russian)

[5] Bythovskiĭ, M. L., "The accuracy of electric networks intended for the solution of Laplace's equation," *Izvestiya* Akad. Nauk SSSR. Otd. Tehn. Nauk **1950**, 1950, 489–526.

[6] Bythovskiĭ, M. L., "The accuracy of electrical circuits for calculation," *Izvestiya* Akad. Nauk SSSR. Otd. Tehn. Nauk **1948**, 1948, 1239–1278.

[7] Chadaja, F. G., "On the error in the numerical integration of ordinary differential equations by the method of finite differences," *Trav.* Inst. Math. Tbilissi [Trudy Tbiliss.Mat. Inst.] **11**, 1942, 97–108.

[8] Coddington, E. A. and Levinson, N., "A boundary value problem for a non-linear

differential equation with small parameter," *Proc.* of the A. M. S., **3**, No. 1, 1952, 73–81.

[9] COLLATZ, L. AND ZURMÜHL, R., "Beiträge zu den Interpolationsverfahren der numerischen Integration von Differentialgleichungen 1. and 2. Ordnung," *Z. Angew. Math. Mech.* **22**, 1942, 42–55.

[10] DUNCAN, W. J., "Assessment of errors in approximate solutions of differential equations," *Coll. Aeronaut. Cranfield. Rep.* no. **13**, 1947, 9 pp.

[11] DUNCAN, W. J., "Assessment of errors in approximate solutions of differential equations," *Quart. J. Mech. Appl. Math.*, **1**, 1948, 470–476.

[12] DUNCAN, W. J., "Technique of the step-by-step integration of ordinary differential equations," *Philos. Mag.* (7), **39**, 1948, 493–509.

[13] FORSYTHE, G. E., "Note on rounding-off errors," National Bureau of Standards, Los Angeles, Calif. 1950, 3 pp.

[14] FRICKE, A., "Über die Fehlerabschätzung des Adamsschen Verfahrens zur Integration gewöhnlicher Differentialgleichungen 1. Ordnung," *Z. Angew. Math. Mech.*, **29**, 1949, 165–178.

[15] FRIEDRICHS, K. O. AND WASOW, W. R., "Singular perturbations of non-linear oscillations," *Duke Math. Jn.*, **13**, 1946, 367–381.

[16] GILL, S., "A process for the step-by-step integration of differential equations in an automatic digital computing machine," *Proc.*, Cambridge Phil. Soc., **47**, 1951, 96–108.

[17] GRADŠTEĬN, I. S., "Linear equations with variable coefficients and small parameters in the highest derivatives," Mat. Sbornik N. S., **27**, (**69**), 1950, 47–68.

[18] HUSKEY, H. D., "On the precision of a certain procedure of numerical integration," *Jn. Research*, Nat. Bur. Standards, **42**, 1949, 57–62.

[19] KIRBY, S., "The relative accuracy of quadrature formulae of the Cotes' closed type," *Coll. Aeronaut. Cranfield. Rep.* no. **17**, 1948, 6 pp.

[20] KOBRINSKIĬ, N. E. AND LYUSTERNIK, L. A., "Mathematical Technics," Uspehi Matem. Nauk (N. S.) **1**, no. 5–6, (15–16), 1946, 3–26.

[21] KORN, G. A., "The difference analyzer: A simple differential equation solver," *Math. Tables and Other Aids to Comput.*, **6**, no. 34, 1952, 1–8.

[22] KORN, G. A., "Elements of D. C. analog computers," *Electronics*, **21**, no. 4, 1948, 122–127.

[23] KORN, G. A., "Design of D. C. electronic integrators," *Electronics*, **21**, no. 5, 1948, 124–126.

[24] LAHAYE, E., "Une méthode de résolution des équations différentielles," Acad. Roy. Belgique *Bull.* Cl. Sci. (5) **34**, 1948, 851–862.

[25] LEVINSON, N., "Perturbations of discontinuous solutions of non-linear systems of differential equations," *Acta* Mathematica, **82**, 1947, 71–76.

[26] LOUD, W. S., "On the long-run error in the numerical solution of certain differential equations," *Jn. Math. Phys.*, **28**, 1949, 45–49.

[27] MACNEE, A. B., "An electronic differential analyzer," *Proc.* I. R. E., **37**, 1949, 1315–1324.

[28] MACNEE, A. B., "Some limitations on the accuracy of electronic differential analyzers," *Proc.* I. R. E., **40**, 1952, 303–308.

[29] MILLER, J. C. P., "Checking by differences. I," *Math. Tables and Other Aids to Comput.*, **4**, 1950, 3–11.

[30] MURRAY, F. J., "Planning and error considerations for the numerical solution of a system of differential equations on a sequence calculator," *Math. Tables and Other Aids to Comput.*, **4**, 1950, 133–144.

[31] MURRAY, F. J., "Error Analysis for mathematical machines," *Trans.* of the N. Y. Acad. Sci., **13**, 1951, 168–174.

[31A] MURRAY, F. J., "Mathematical Error Analysis for Continuous Computers," *Project Cyclone Symposium II*, Part II, Reeves Instrument Corporation, U. S. Navy

Special Devices Center and U. S. Navy Bureau of Aeronautics, 1952, 139–146, New York.

[32] PAPOULIS, A., "On the accumulation of errors in the numerical solution of differential equations," *Jn. Appl. Phys.*, **23**, 1952, 173–176.

[33] RADEMACHER, H. A., "On the accumulation of errors in processes of integration on high-speed calculating machines," *Proc.*, Symposium on large scale digital calculating machinery, The Annals of the Computation Laboratory of Harvard University, **16**, Harvard University Press, 1948, 176–187.

[34] RAYMOND, F. H., "Sur un type général de machines mathematiques algebriques," *Ann. Telecommun.*, **5**, 1950, 2–20.

[35] RICE, S. O., "Mathematical analysis of random noise," *Bell Sys. Tech. Jn.*, **23**, no. 3, 1944, 282–332.

[36] TODD, J., "Notes on modern numerical analysis. I," *Math. Tables and Other Aids to Comput.*, **4**, 1950, 39–44.

[37] TIHONOV, A. N., "On systems of differential equations containing parameters," Mat. Sbornik N. S., **27**, (**69**), 1950, 147–156.

[38] TURTON, F. J., "The errors in the numerical solution of differential equations," *Philos. Mag.* **28**, 1939, 359–363.

[39] TURTON, F. J., "Two notes on the numerical solution of differential equations," *Philos. Mag.* **28**, 1939, 381–384.

[40] VASIL'EVA, A. B., "On differentiation of solutions of systems of differential equations containing a small parameter," *Doklady* Akad. Nauk SSSR (N. S.) **75**, 1950, 483–486.

[41] VASIL'EVA, A. B., "On the differentiation of solutions of differential equations containing a small parameter," *Doklady* Akad. Nauk SSSR (N. S.) **61**, 1948, 597–599.

NEW YORK UNIVERSITY
COLUMBIA UNIVERSITY

(Received June 6, 1952; Revised October 28, 1952)

NOTE ON S. N. LIN'S METHOD OF FACTORING POLYNOMIALS

By F. B. Hildebrand

1. Introduction. An iterative numerical procedure for obtaining roots of algebraic equations was devised in 1941 by S. N. Lin [1]. The method consists basically in reducing a polynomial to the product of two factors, and has been extended by Luke and Ufford [2] to the general case in which more than two factors may be obtained in a single sequence of iterations. Because of the simplicity and compactness of the relevant computational program, the method has found considerable favor. However, there exists no satisfactory criterion for predicting convergence or divergence of the iteration sequence.

The present note deals with the special case in which a single quadratic factor is to be extracted, and obtains a precise criterion for asymptotic *stability* of the iteration. In the case of a quartic, the result is capable of a very simple geometric interpretation. Whereas conditions for actual *convergence* are not established, the results obtained serve to display certain curious features of the method.

2. The iterative process. If a polynomial $f(z)$, of degree n, is divided by a quadratic $z^2 + pz + q$, the result can be written in the form

$$\begin{aligned} f(z) &\equiv z^n + a_1 z^{n-1} + \cdots + a_n \\ &= (z^2 + pz + q)(z^{n-2} + b_1 z^{n-3} + \cdots + b_{n-2}) + r_0 z + r_1, \end{aligned} \tag{1}$$

where the recurrence formula

$$\left.\begin{aligned} b_r + pb_{r-1} + qb_{r-2} &= a_r \qquad (r = 1, 2, \cdots, n) \\ b_{-1} &= 0, \qquad b_0 = 1 \end{aligned}\right\} \tag{2}$$

defines the coefficients of the quotient as functions of p and q for $0 \leqq r \leqq n - 2$. Also, the coefficients of the remainder are given by

$$r_0 = a_{n-1} - pb_{n-2} - qb_{n-3} = b_{n-1} \tag{3a}$$

and

$$r_1 = a_n - qb_{n-2} = b_n + pb_{n-1}. \tag{3b}$$

Lin's iteration defines the coefficients of the modified quadratic $z^2 + p'z + q'$ by the equations

$$q' = \frac{a_n}{b_{n-2}}, \qquad p' = \frac{a_{n-1} - qb_{n-3}}{b_{n-2}}, \tag{4a, b}$$

and hence consists in solving the equations $r_0(p, q) = 0$ and $r_1(p, q) = 0$ by successive substitutions. For present purposes, it is more convenient to write these equations in the equivalent form

$$\frac{p' - p}{q'} = \frac{b_{n-1}}{a_n} \equiv \frac{r_0}{a_n}, \qquad \frac{q' - q}{q'} = \frac{b_n + pb_{n-1}}{a_n} \equiv \frac{r_1}{a_n}. \tag{5a, b}$$

The explicit solution of (2) is obtained in the form

$$b_r = \frac{1}{\beta - \alpha} \sum_{s=0}^{r} a_s(\beta^{r-s+1} - \alpha^{r-s+1}) \tag{6}$$

where α and β are the zeros of $z^2 + pz + q$, so that

$$p = -(\alpha + \beta), \qquad q = \alpha\beta. \tag{7}$$

In particular, there follows

$$r_0 \equiv b_{n-1} = \frac{1}{\beta - \alpha}[f(\beta) - f(\alpha)], \tag{8a}$$

$$r_1 \equiv b_n + pb_{n-1} = \frac{1}{\beta - \alpha}[\beta f(\alpha) - \alpha f(\beta)], \tag{8b}$$

so that equations (5a, b) take the form

$$\frac{p' - p}{q'} = \frac{f(\beta) - f(\alpha)}{a_n(\beta - \alpha)}, \qquad \frac{q' - q}{q'} = \frac{\beta f(\alpha) - \alpha f(\beta)}{a_n(\beta - \alpha)} \tag{9a, b}$$

where α and β satisfy (7).

3. Asymptotic stability. It is now assumed that, for (p, q) sufficiently near to (P, Q), where P and Q are the true coefficients of the quadratic factor relevant to the roots z_1 and z_2,

$$P = -(z_1 + z_2), \qquad Q = z_1 z_2, \tag{10}$$

it is permissible to linearize equations (9a, b) in the errors $(p - P)$ and $(q - Q)$.

From equations (7) and (8) it follows, after a simple calculation, that when $(p, q) = (P, Q)$, and hence $(\alpha, \beta) = (z_1, z_2)$, one has

$$\begin{aligned} \frac{\partial r_0}{\partial p} &= -\frac{z_1 f'(z_1) + z_2 f'(z_2)}{(z_2 - z_1)^2} \equiv c_{11}, & \frac{\partial r_0}{\partial q} &= -\frac{f'(z_1) + f'(z_2)}{(z_2 - z_1)^2} \equiv c_{12}, \\ \frac{\partial r_1}{\partial p} &= z_1 z_2 \frac{f'(z_1) + f'(z_2)}{(z_2 - z_1)^2} \equiv c_{21}, & \frac{\partial r_1}{\partial q} &= \frac{z_2 f'(z_1) + z_1 f'(z_2)}{(z_2 - z_1)^2} \equiv c_{22}. \end{aligned} \tag{11}$$

Thus the result of linearizing (9a, b) can be written in the form

$$u_{k+1} = u_k + \gamma(c_{11}u_k + c_{12}v_k), \qquad v_{k+1} = v_k + \gamma(c_{21}u_k + c_{22}v_k), \tag{12}$$

with the abbreviation

$$\gamma = z_1 z_2 / a_n \tag{13}$$

where u_k and v_k represent the errors $p - P$ and $q - Q$, respectively, in the kth iteration.

The errors u_k and v_k are accordingly linear combinations of λ_1^k and λ_2^k, where λ_1 and λ_2 are the latent roots of the matrix

$$\begin{bmatrix} 1 + \gamma c_{11} & \gamma c_{12} \\ \gamma c_{21} & 1 + \gamma c_{22} \end{bmatrix}, \tag{14}$$

and are obtainable in the form

$$\lambda_1 = 1 + \frac{z_1 z_2}{z_2 - z_1}\frac{f'(z_1)}{f(0)}, \qquad \lambda_2 = 1 - \frac{z_1 z_2}{z_2 - z_1}\frac{f'(z_1)}{f(0)}. \tag{15}$$

It follows that *Lin's iteration is asymptotically stable near the root pair* (z_1, z_2) *if and only if* $|\lambda_1| < 1$ *and* $|\lambda_2| < 1$, *or if*

$$\left|1 - \left(1 - \frac{z_i}{z_3}\right)\left(1 - \frac{z_i}{z_4}\right)\cdots\left(1 - \frac{z_i}{z_n}\right)\right| < 1 \qquad (i = 1, 2) \tag{16}$$

where $z_3, \cdots, z_n$ *are the remaining roots of the equation* $f(z) = 0$.

This result complements one obtained by Friedman [3], in the case when Lin's iteration is applied to the extraction of a *linear* factor, the condition for asymptotic stability near a single root z_1 being of the form

$$\left|1 + z_1\frac{f'(z_1)}{f(0)}\right| < 1 \tag{17}$$

or, equivalently,

$$\left|1 - \left(1 - \frac{z_1}{z_2}\right)\left(1 - \frac{z_1}{z_3}\right)\cdots\left(1 - \frac{z_1}{z_n}\right)\right| < 1. \tag{18}$$

4. Specialization to the quartic. If $f(z)$ is of degree four, the conditions (16) take the form

$$|z_i|\,|z_i - (z_3 + z_4)| < |z_3 z_4| \qquad (i = 1, 2), \tag{19}$$

and are subject to an interesting geometric interpretation.

In the complex plane, the locus

$$|z|\,|z - 2a| = c^2, \tag{20}$$

where c is real, is a so-called *Cassinian* (that is, the locus of points such that the product of their distances from two fixed foci, at $z = 0$ and at $z = 2a$, is constantly c^2). If $e \equiv |a|/c$, the locus comprises two ovals when $e > 1$, a lemniscate when $e = 1$, a doubly indented oval when $\frac{1}{2}\sqrt{2} < e < 1$, a simple (unindented) oval when $0 < e \leqq \frac{1}{2}\sqrt{2}$, and a circle when $e = 0$.

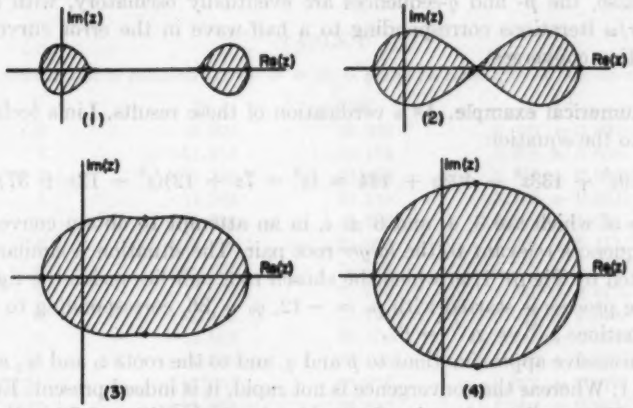
Hence, if the *Cassinian* of the root pair (z_3, z_4) is defined as the locus of points such that the product of the distances from the origin and from the point $z = z_3 + z_4$ is equal to $|z_3 z_4|$, it follows from (19) that *the Lin iteration is asymptotically stable near a root pair* (z_1, z_2) *if and only if that root pair lies inside the Cassinian of the complementary pair* (z_3, z_4).

The Cassinian of (z_3, z_4) passes through the points $z = z_3$ and $z = z_4$, with its center at $z = \frac{1}{2}(z_3 + z_4)$ and foci at the origin and at $z = z_3 + z_4$, and the "eccentricity" is given by

$$e = \frac{|a|}{c} = \frac{|z_3 + z_4|}{2\sqrt{|z_3 z_4|}}.$$

In particular, when z_3 and z_4 are real and of common sign, they lie at the

inner intercepts of a pair of ovals (Fig. 1: open points) or at the center of a lemniscate (Fig. 2) when they coincide. When z_3 and z_4 are real and of opposite sign, four cases exist. If z_3 is nearest the origin, the points lie at the outer intercepts of a pair of ovals (Fig. 1: solid points) when $|z_3|/|z_4| < 3 - 2\sqrt{2}$, at the extremities of a lemniscate (Fig. 2) when $|z_3| = (3 - 2\sqrt{2})|z_4|$, at the extremities of an indented oval (Fig. 3: solid points) when $(3 - 2\sqrt{2}) < |z_3|/|z_4| < \frac{1}{2}(3 - \sqrt{5})$, and at the extremities of a simple oval (Fig. 4: solid points) otherwise.



FIGURES 1 to 4. Typical Stable Root-Pair Areas in Lin Iteration for Quartic Equations.

If z_3 and z_4 are conjugate complex, they lie at the ends of the central transversal of an oval which is indented when $|Re(z_3)| > |Im(z_3)|$ and simple otherwise (Figs. 3 and 4: open points). The axial span is then $2\{[Re(z_3)]^2 + |z_3|^2\}^{\frac{1}{2}}$, the transversal at the center $2|Im(z_3)|$. When the oval is indented, the *maximum* transversal is of length $|z_3|^2/|Re(z_3)|$, at axial distance $\{4[Re(z_3)]^4 - |z_3|^4\}^{\frac{1}{2}}/[2|Re(z_3)|]$ from the center. If $z_{3,4} = \rho e^{\pm i\phi}$, the inscribed *circle* with center at the origin is of radius $\rho(\sqrt{1 + \cos^2\phi} - |\cos\phi|)$.

It is seen that, for a quartic, the smaller pair of roots is not necessarily stable; *nor is the larger pair necessarily unstable.* If there are two pairs of conjugate complex roots $z_{1,2} = re^{\pm i\theta}$ and $z_{3,4} = \rho e^{\pm i\phi}$, the preceding result shows that a *sufficient* (but quite conservative) condition for stability of (z_1, z_2) is the requirement $r/\rho < \sqrt{2} - 1 \doteq 0.41$.

Since the asymptotic expressions for the errors in p and q in the kth cycle are each of the form $c_1\lambda_1^k + c_2\lambda_2^k$, where c_1 and c_2 are independent of k, it follows that, when the iteration converges, the limiting *error damping factors* are $|\lambda_1|$ and $|\lambda_2|$, where

$$\lambda_i = \frac{z_i}{z_3 z_4}[(z_3 + z_4) - z_i] = 1 - \left(1 - \frac{z_i}{z_3}\right)\left(1 - \frac{z_i}{z_4}\right)$$
$$= \frac{1}{z_3 z_4}\{z_1 z_2 - [(z_1 + z_2) - (z_3 + z_4)]z_i\} \qquad (i = 1, 2). \qquad (21)$$

The sequence of successive approximations to p and q is eventually nonoscillatory if λ_1 and λ_2 are real and non-negative. If (z_1, z_2) are conjugate complex and the coefficients of the polynomial are real, the asymptotic expressions for the errors in p and q are each of the form $A\,|\,\lambda_1\,|^k \sin(\omega k + B)$, where A and B are real constants, and where

$$\cos\omega = \frac{Re(\lambda_1)}{|\,\lambda_1\,|}. \tag{22}$$

In this case, the p- and q-sequences are eventually oscillatory, with approximately π/ω iterations corresponding to a half-wave in the error curves, when the iteration converges.

5. A numerical example. As a verification of these results, Lin's technique is applied to the equation

$$z^4 - 19z^3 + 133z^2 - 403z + 444 \equiv (z^2 - 7z + 12)(z^2 - 12z + 37) = 0,$$

the roots of which are 3, 4, and $6 \pm i$, in an attempt to obtain convergent p- and q-sequences relevant to the *larger* root pair. The situation is similar to that represented by Figure 1, in which the chosen root pair lies inside the right-hand oval. The process is started with $p_0 = -12$, $q_0 = 36$, corresponding to the root approximations $z_1^{(0)} = z_2^{(0)} = 6$.

The successive approximations to p and q, and to the roots z_1 and z_2, are listed in Table 1. Whereas the convergence is not rapid, it is indeed present. Equations (21) and (22) predict a damping factor $|\,\lambda_{1,2}\,| = \sqrt{74}/12 \doteq 0.72$ in the amplitudes of the errors in p and q, and an oscillation with half-wave corresponding to $\pi/\omega = \pi/\tan^{-1}\frac{5}{7} \doteq 5$ iterations (near the end of the sequence of calculations). The tabulated results are in accordance with these predictions, the relatively slow rate of convergence corresponding to the relative nearness of $|\,\lambda\,|$ to unity.

In spite of the oscillation of the p- and q-sequences, it may be verified that, in this case, the actual distance between a true root and its kth approximation decreases *steadily* as k increases, the points representing successive approximations describing a spiral path about the true root in the complex plane.

6. Monotonic root convergence. The eventual steady approach of root-pair approximations to roots, noted in the preceding example, exists, indeed, for *any* algebraic equation, when the Lin iteration converges. In fact, if equations (9a, b) are linearized in the errors $\alpha - z_1$ and $\beta - z_2$, the results are found to be equivalent to the equations

$$\alpha_{k+1} - z_1 = \lambda_1(\alpha_k - z_1), \qquad \beta_{k+1} - z_2 = \lambda_2(\beta_k - z_2),$$

from which it follows that, asymptotically,

$$\alpha_k - z_1 = A\lambda_1^k, \qquad \beta_k - z_2 \doteq B\lambda_2^k, \tag{23a, b}$$

where A and B are independent of k. Hence $|\,\alpha_k - z_1\,|$ and $|\,\beta_k - z_2\,|$ eventually decrease steadily in cases of convergence.

Equations (23a, b) can be used in an obvious way to obtain a more accurate determination of z_1 and z_2 from the results of three successive convergent iterations, if sufficiently many significant figures are retained in the calculations.

The absence of coupling between (23a) and (23b) suggests that, if only one member of a root pair, say z_1, is stable, whereas the other root z_2 is unstable (so that $|\lambda_1| < 1$ and $|\lambda_2| > 1$), and if α_k and β_k are the zeros of $z^2 + p_k z + q_k$, then α_k may tend to z_1 even though the β-, p-, and q-sequences diverge. This conjecture was tested empirically in the case of the equation $z^4 - 4z^3 + 7z^2 - 16z + 12 = 0$, with the roots $z_1 = 1$, $z_2 = 3$, and $z_{3,4} = \pm 2i$, the iteration being

TABLE 1

Approximations to quadratic factor $z^2 + pz + q$ and its zeros after k Lin iterations

k	p_k	q_k	$z_{1,2}^{(k)}$
0	−12.000	36.000	$6.000 \pm 0.000i$
1	−11.615	34.154	$5.808 \pm 0.651i$
2	−11.536	33.968	$5.768 \pm 0.836i$
3	−11.562	34.346	$5.781 \pm 0.963i$
4	−11.658	35.086	$5.829 \pm 1.054i$
5	−11.801	36.038	$5.901 \pm 1.111i$
6	−11.957	36.979	$5.978 \pm 1.119i$
7	−12.073	37.605	$6.037 \pm 1.077i$
8	−12.112	37.731	$6.056 \pm 1.028i$
9	−12.086	37.500	$6.043 \pm 0.991i$
10	−12.040	37.193	$6.020 \pm 0.977i$
11	−12.003	36.978	$6.002 \pm 0.979i$
12	−11.984	36.880	$5.992 \pm 0.987i$
13	−11.980	36.872	$5.990 \pm 0.995i$
14	−11.985	36.910	$5.992 \pm 1.002i$
.....	.....	.....	.....
∞	−12.000	37.000	$6.000 \pm 1.000i$

started with $p_0 = q_0 = 0$. Here $\lambda_1 = \frac{1}{4}$, $\lambda_2 = \frac{9}{4}$, and $|\lambda_{3,4}| = \frac{4}{3}\sqrt{5}$, so that while z_1 is stable, no stable root *pairs* exist. It was found that, whereas the p- and q-sequences appear to diverge, one zero of $z^2 + p_k z + q_k$ does indeed tend rapidly to unity.

7. Remarks. In his original paper [1], Lin also suggested a modified procedure in which q is replaced by q' in (4b). Whereas stability of this procedure has not been investigated completely,* numerical calculation indicates that the modified procedure does not converge to the coefficients of the quadratic relevant to the chosen root pair in the case of the present example (Section 5), but that it converges instead to the coefficients corresponding to the pair (3, 4).

Another method, due to Friedman [3], is somewhat similar to Lin's method, but it involves about twice as much calculation per iteration. In the special case of the determination of the smallest zero of a polynomial with only real

* In the case when $(z - z_1)(z - z_2)$ is to be extracted from a *cubic*, the stability condition can be shown to be $|1 - (1 - \xi_1)(1 - \xi_2)| < 1$ where $\xi_i = z_i/z_3 (i = 1, 2)$, whereas (16) reduces to the two conditions $|\xi_i| < 1$ $(i = 1, 2)$.

zeros of common sign, both methods are stable, but Friedman's method has been shown to have a faster asymptotic rate of error decay. In other cases, either may be stable while the other is not. Limited information is available [3] concerning the stability of Friedman's method when a quadratic factor is to be extracted. It appears to yield a divergent sequence in the present example.

There exists also a method due to Bairstow [4], rediscovered by Hitchcock [5] and also systematized by Milne [6], which differs from Lin's method in that the equations $r_0(p, q) = r_1(p, q) = 0$ are treated by Newton-Raphson iteration. The calculations are similar to those of the Lin technique, but each iteration again involves about twice as much labor. In the present example, it gives a more rapidly convergent sequence of approximations (leading to the quadratic factor associated with the larger root-pair) than does the Lin method. Since, in a convergent Newton-Raphson sequence, the errors in the $(k + 1)$th iteration generally are asymptotically linear combinations of the *squared* errors in the kth iteration, a similarly favorable comparison with the methods of Lin and Friedman may be expected, more generally, whenever the Bairstow sequence is convergent.

It may be remarked, in conclusion, that the stability criterion obtained in this note is clearly of limited practical significance unless it is feasible to obtain fair approximations to the roots by other methods before the iteration is initiated. Moreover, the problem of determining those regions in the complex plane inside which the zeros of the initially assumed quadratic must lie, in order that the iteration converge to a stable pair of zeros of $f(z)$, remains to be studied. In any case, the results obtained may serve to emphasize the fact that "practical rules" which are stated in terms of the *magnitudes* relevant to the degree of "root separation," and which disregard its vectorial nature, are either highly conservative or undependable, when applied to the Lin iteration.

REFERENCES

1. S. N. Lin: A Method of Successive Approximations of Evaluating the Real and Complex Roots of Cubic and Higher Order Equations. Journ. Math. and Phys. **20**, 231–242 (1941).
2. Y. L. Luke and D. Ufford: On the Roots of Algebraic Equations. Journ. Math. and Phys. **30**, 94–101 (1951).
3. B. Friedman, Note on Approximating Complex Zeros of a Polynomial. Comm. Pure and Appl. Math. **2**, 195–208 (1949).
4. L. Bairstow, Reports and Memoranda No. 154, Advisory Committee for Aeronautics (1914).
5. F. L. Hitchcock, An Improvement on the G. C. D. Method for Complex Roots. Journ. Math. and Phys. **23**, 69–74 (1944).
6. W. E. Milne, *Numerical Calculus*. Princeton University Press, Princeton (1949), pp. 53–57.

Massachusetts Institute of Technology

(Received October 20, 1952)

A NEW INTEGRATION PROCEDURE

By M. Lotkin

1. Introduction. In a recently published article [1] there was described a highly accurate procedure for the integration of ordinary differential equations. This procedure, being of the Euler-Maclaurin type, necessitates the knowledge of first and second order derivatives at the endpoints of the interval of integration.[1]

Now in practice it is a frequent occurrence that the functions whose derivatives must be computed are known only at certain prescribed, equidistant arguments, so that the derivatives, if they exist at all, are not immediately available. It is thus of interest to extend the range of applicability of the above mentioned process to functions given in tabular form.

This extension has been carried out, and is described in detail in the following sections. Several new integration formulas are obtained, their accuracy is stated and illustrated by means of appropriate examples. The suitability of these formulas for high speed computing machinery is obvious.

The new formulas derived differ essentially from other quadrature formulas in two important respects. First, the new formulas apply to the basic interval of length h, thus permitting quadratures over ranges of arbitrary length to be obtained quite simply. Second, they differ from open-type formulas, which employ interior points only, and closed-type formulas, which employ interior as well as terminal points, in that they make use also of exterior points, thus tending to produce frequently smoother and more accurate results.

A generalization of the method described in [1] to the solution of differential equations will be dealt with subsequently.

2. The quadrature formulas. The basic formula of reference [1] is

$$I_1 \equiv \int_{x_0}^{x_1} f(x)\,dx = h[\tfrac{1}{2}(f_0 + f_1) + \tfrac{1}{10}h(f_0' - f_1') + \tfrac{1}{120}h^2(f_0'' + f_1'')] + R, \tag{2.1}$$

with $h = x_1 - x_0$, $f_i = f(x_i)$, $f_j^{(k)} = (d^k f/dx^k)_{x=x_j}$, $j = 0, 1$, $k = 1, 2$, and $R = -(h^7/100{,}800)\,f^{vi}(\xi)$, $x_0 < \xi < x_1$.

Let us assume now that the $f_i = f(x_i)$ are known for a succession of equidistant arguments x_i, $i = 0, \pm 1, \pm 2, \cdots$ Then the derivatives $f_i^{(k)}$ in (2.1) may be computed by means of central differences [2] as follows:

$$\begin{aligned} hf_i' &= \mu\,\delta f_i - \tfrac{1}{6}\mu\,\delta^3 f_i + \tfrac{1}{30}\mu\,\delta^5 f_i - \tfrac{1}{140}\mu\,\delta^7 f_i + \cdots \\ h^2 f_i'' &= \delta^2 f_i - \tfrac{1}{12}\delta^4 f_i + \tfrac{1}{90}\delta^6 f_i - \tfrac{1}{560}\delta^8 f_i + \cdots \end{aligned} \tag{2.2}$$

[1] It has come to my attention recently that the procedure employed in [1] was previously described by W. E. Milne, in [6].

These differences $\delta^k f_i$, and means $\mu\delta^k f_i$, are defined thus:

$$\delta^{k+1} f_i = \delta^k f_{i+\frac{1}{2}} - \delta^k f_{i-\frac{1}{2}}$$

$$\mu\delta^k f_i = \tfrac{1}{2}(\delta^k f_{i+\frac{1}{2}} + \delta^k f_{i-\frac{1}{2}}),$$

for $k = 0, 1, 2, \cdots$, with $\delta^0 f = f$.

A short calculation shows that

$$\delta^{2k} f_i = \sum_{r=0}^{2k} (-1)^r \binom{2k}{r} f_{i+k-r}$$

$$\mu\delta^{2k+1} f_i = \tfrac{1}{2}[(f_{i+1+k} - f_{i-1-k}) - 2k(f_{i+k} - f_{i-k})] + \sum_{r=2}^{k} (-1)^r C_{kr}(f_{i+1+k-r} - f_{i-1-k+r})$$

with

$$C_{kr} = \binom{2k}{r} - \binom{2k}{r-2}.$$

In stopping in the series (2.2) after a certain term there will be introduced into (2.1) a new truncation error whose magnitude will depend on the first term neglected. Thus, taking one, two, or three terms in (2.2) results in new formulas denoted as "three-, five,- or seven-point" quadrature rules. These turn out to be quite useful for practical purposes, and will now be discussed separately. It will be seen that the extension of these quadratures formulas to regions comprising several successive intervals leads to certain general combinations of the functional entries f_i of the form

$$F_{-m,s+m} = (\sum_{-m}^{m} + 2\sum_{m+1}^{s-m-1} + \sum_{s-m}^{s+m}) f_i \,. \tag{2.3}$$

Thus, e.g.,

$$F_{0,s} = f_0 + 2\sum_{i=1}^{s-1} f_i + f_s$$

$$F_{-1,s+1} = (f_{-1} + f_0 + f_1) + 2\sum_{2}^{s-2} f_i + (f_{s-1} + f_s + f_{s+1}), \text{ etc.}$$

A. The Three-Point Formula. From (2.2) we have

$$hf_i' = \mu\,\delta f_i - \tfrac{1}{6}h^3 f'''(\xi_i')$$

$$h^2 f_i'' = \delta^2 f_i - \tfrac{1}{12}h^4 f^{iv}(\xi_i''),\ x_{i-1} < \xi_i',\ \xi_i'' < x_{i+1}\,.$$

Consequently,

$$I_1 = h[\tfrac{1}{2}(f_0 + f_1) + \tfrac{1}{10}h(f_0' - f_1') + \tfrac{1}{120}h^2(f_0'' + f_1'')] + R$$

$$= \tfrac{1}{24}h[13(f_1 + f_0) - (f_2 + f_{-1})] + R_3\,,$$

$$R_3 = \tfrac{1}{60}h^4[f'''(\xi_1') - f'''(\xi_0')] - \tfrac{1}{1440}h^5[f^{iv}(\xi_0'') + f^{iv}(\xi_1'')] + R.$$

We may then conclude that

$$I_1 = \tfrac{1}{24}h[13(f_1 + f_0) - (f_2 + f_{-1})] + R_3\,,$$

$$R_3 = C_3(h^5/5!)\, f^{iv}(\xi_3),\ x_{-1} < \xi_3 < x_2. \tag{2.4}$$

For functions f that are of the polynomial type of degree not exceeding four, C_3 is 11/6. This formula is thus exact for polynomials $f(x)$ of third or lower degree. As the remainder term indicates the accuracy of (2.4) is approximately that of Simpson's Rule

$$I_2 = \tfrac{1}{3}h[f_0 + 4f_1 + f_2] - \tfrac{1}{90}h^5 f^{iv}(\xi).$$

Formula (2.4) is readily extended to regions comprising several successive intervals:

$$I_s \equiv \int_{x_0}^{x_s} f(x)\,dx = \tfrac{1}{24}h[13F_{0,s} - F_{-1,s+1}] + C_3\,(h^5/5!)\sum_{i=1}^{s} f^{iv}(\xi_{3i}),\ x_{i-2} < \xi_{3i} < x_{i+1}, \tag{2.5}$$

with $F_{0,s}, F_{-1,s+1}$ defined in (2.3).

The fact that

$$R_3 = 11\tfrac{1}{720}h^5 f^{iv}(\xi_3), \qquad x_{-1} < \xi_3 < x_2,$$

has a sign generally opposite to that of

$$R_s = -\tfrac{1}{90}h^5 f^{iv}(\xi), \qquad x_0 < \xi < x_2$$

may be used to derive a formula of slightly higher accuracy than either (2.4) or Simpson's rule, as follows:

Putting

$$G_i = \tfrac{1}{24}h[13(f_i + f_{i-1}) - (f_{i-3} + f_{i-2})]$$

$$S_i = \tfrac{1}{24}h[32f_i + 8(f_{i+1} + f_{i-1})]$$

we have, on one hand,

$$I_2 = G_1 + G_2 + R_{31} + R_{32},$$

and, on the other,

$$I_2 = S_1 + R_{s1}.$$

Thus

$$I_2 = \tfrac{1}{2}(G_1 + G_2 + S_1) + \tfrac{7}{720}h^5 f^{iv}(\bar{\xi}), \qquad x_{-1} < \bar{\xi} < x_2.$$

In terms of functional entries, then,

$$I_2 \approx \tfrac{1}{24}h[29f_1 + 10(f_2 + f_0) - \tfrac{1}{2}(f_3 + f_{-1})]. \tag{2.5a}$$

B. The Five-Point Formula. In this case we find that

$$hf_i' = \mu\,\delta f_i - \tfrac{1}{6}\mu\,\delta^3 f_i + \tfrac{1}{30}h^5 f^{v}(\xi_i')$$

$$h^2 f_i'' = \delta^2 f_i - \tfrac{1}{12}\delta^4 f_i + \tfrac{1}{90}h^6 f^{vi}(\xi_i''),$$

with $x_{i-2} < \xi_i', \xi_i'' < x_{i+2}$. It turns out that

$$I_1 = \tfrac{1}{1440}h[802(f_1 + f_0) - 93(f_2 + f_{-1}) + 11(f_3 + f_{-2})] + R_5,$$
$$R_5 = C_5(h^7/7!)f^{v'}(\xi_5). \tag{2.6}$$

Here $C_5 = -191/12$ for functions f that are polynomials of degree not exceeding six.

Applying (2.6) to s successive intervals we obtain

$$I_s = \tfrac{1}{1440}h[802F_{0,s} - 93F_{-1,s+1} + 11F_{-2,s+2}] + C_5(h^7/7!)\sum_{i=1}^{s} f^{v'}(\xi_{5i}) \tag{2.7}$$

For $s = 6$, then, the truncation error is approximately $6C_5(h^7/7!)f^{v'}(\xi_5)$, which is of the same order of magnitude as the remainder in Weddle's rule:

$$I_6 = \tfrac{3}{10}h[(f_0 + f_6) + 5(f_1 + f_5) + (f_2 + f_4) + 6f_3] + R_w,$$
$$R_w = -\tfrac{1}{140}h^7 f^{v'}(\xi).$$

C. The Seven-Point Formula. Taking three terms in the series (2.2) we get the remainders $R' = -\tfrac{1}{140}h^7 f^{v'}(\xi_i')$, $R'' = -\tfrac{1}{560}h^8 f^{v'''}(\xi_i'')$, with the ξ_i', ξ_i'' located in (x_{i-3}, x_{i+3}). The total truncation error in the quadrature formula (2.1) is now of the form $R_7 = R + C(h^9/1400)f^{v'''}(\xi_7)$, with $R = -(\tfrac{1}{100.800}h^7 f^{v'}(\xi), x_0 < \xi < x_1$, denoting the remainder term of (2.1). In this case, then, the accuracy of (2.1) is preserved if h is taken sufficiently small.

In replacing f_i', f_i'' by means of (2.2) it is found that

$$hf_i' = \tfrac{1}{60}[45(f_{i+1} - f_{i-1}) - 9(f_{i+2} - f_{i-2}) + (f_{i+3} - f_{i-3})] + R'$$
$$h^2 f_i'' = \tfrac{1}{180}[270(f_{i+1} + f_{i-1}) - 27(f_{i+2} + f_{i-2}) + 2(f_{i+3} + f_{i-3}) - 490f_i] + R''.$$

It follows that

$$I_1 = \tfrac{1}{21.600}h[12{,}200\,(f_1 + f_0) - 1701(f_2 + f_{-1}) + 335(f_3 + f_{-2}) - 34(f_4 + f_{-3})] + R_7 \tag{2.8}$$

For s successive intervals, then,

$$I_s = \tfrac{1}{21.600}h[12{,}200F_{0,s} - 1701F_{-1,s+1} + 335F_{-2,s+2} - 34F_{-3,s+3}] + \sum_i R_{7i} \tag{2.9}$$

For sufficiently small h, $R_7 \approx R$, and then $6\,|\,R_7\,|\,/\,|\,R_w\,| \approx \tfrac{1}{120}$, so that (2.9) may be more accurate than Weddle's rule by as many as two decimals.

G. Blanch and I. Rhodes [3] have compounded similar quadrature formulas based on seven-point Lagrangian polynomials. Their formula (8) is:

$$I_s = \tfrac{1}{120.960}h\sum_{i=0}^{6} A_i(f_i + f_{s-i-1}) + h\sum_{j=7}^{s-8} f_j + R_L,$$
$$A_i = 36799,\ 176648,\ 54851,\ 177984,\ 89437,\ 130936,\ 119585,\ |\,R_L\,| < 2 \times 10^{-2}s \cdot h^9 \times |\,f^{v'''}(\xi)\,|,\ s \geqq 14.$$

In those cases which (2.8) is at least as accurate as this formula (8)—and such cases are readily constructed—(2.8) would seem to be preferable, due to its fewer and smaller coefficients.

3. Generalization. The basic formula (2.1) can be extended to higher orders of h, either directly, or through the application of Hummel-Seebeck's generalization of Taylor's expansion [4.][2]

It turns out that

$$I_1 = \sum_{k=1}^{n} D_{nk}[f^{(k-1)}(x_0) - (-1)^k f^{(k-1)}(x_1)]h^k + R_n \tag{3.1}$$

with

$$D_{n,k+1} = \frac{n-k}{(2n-k)(k+1)} D_{nk}, \; D_{n1} = \tfrac{1}{2},$$

$$R_n = (-1^n) \frac{(n!)^2}{(2n)!\,(2n+1)!} h^{2n+1} f^{(2n)}(\xi).$$

In (2.1), $n = 3$. For $n = 2$ formula (3.1) becomes

$$I_1 = \tfrac{1}{2}h(f_0 + f_1) + \tfrac{1}{12}h^2(f_0' - f_1') + h^5 f^{IV}(\xi)/720, \tag{3.2}$$

which is identical with Euler-Maclaurin's formula, terminated after the first derivatives.

For $n = 4$ the coefficients D_{nk} are, respectively, $\frac{1}{2}$, $\frac{3}{28}$, $\frac{1}{84}$, and $\frac{1}{1680}$, and

$$R_4 = h^9 f^{VIII}(\xi)/25,401,600.$$

It is, then, theoretically possible to obtain other sets of multiple-point formulas by replacing the derivatives occurring in (3.1) by central-difference expressions of the type (2.2).

4. Applications. It has been stated many times [5] that in choosing a particular quadrature formula for the evaluation of an integral, or in alleging the superiority of one method over another, a certain amount of caution should be observed. The inherent accuracy of all methods based on polynomial approximation depends essentially on how closely the integrand can be approximated in this manner. It is clear, then, that poor or even absurd values may be obtained if low order polynomials are chosen for integrands that, in or near the interval of integration, are periodic, or asymptotically approach a finite constant, or have diverging differences, or have an infinite slope, etc.

Thus the attainment of sufficient accuracy may necessitate a brief study of the behavior of the integrand, and, if advisable, the partitioning of the total range of integration into suitable subintervals.

The examples described below, illustrating the three formulas previously derived, are, however, sufficiently simple and may be treated in a straightforward manner.

[2] I am indebted to W. M. Stone of Oregon State College for pointing out this connection.

Example 1. Here the three-point formula (2.4) was used to evaluate

$$\int_0^{\pi/2} dx/(1 - \tfrac{1}{2}\sin^2 x)^{\frac{1}{2}}$$

with $h = \pi/12$, carrying ten decimals.

The computed value is shown in the second column of Table 1; it differs from the exact value 1.85407 46773 by an error shown in the third column. The estimated error, computed by means of $\frac{1}{60}h^5 \sum_i f^{IV}(\xi_i)$, is exhibited in the last column.

It is sufficiently close to the actual error to be of practical utility.

TABLE 1

Application of quadrature formulas

	Method	Value	Error	Error Ext.
Ex. 1	Eq. (2.4)	1.85407 46777	5×10^{-10}	2×10^{-7}
	Eq. (2.5a)	1.85407 10984	4×10^{-6}	
	Simpson	1.85406 75191	7×10^{-6}	1×10^{-4}
Ex. 2	Eq. (2.6)	15.70393 8515	5.3×10^{-8}	7.1×10^{-8}
	Weddle	15.70393 8484	2.2×10^{-8}	2.7×10^{-8}
	Steffensen	15.70393 7660	8.0×10^{-7}	
Ex. 3	Eq. (2.9)	0.27272 72723	4×10^{-10}	8×10^{-9}
	Lagrange	0.27272 72751	2×10^{-6}	2×10^{-7}

The value of the integral computed by Simpson's rule happens to be much less accurate than that obtained by the three-point formula. The use of formula (2.5a) also results, as anticipated, in a better value.

Example 2. Next we apply the five-point formula (2.6) to the evaluation of

$$\int_{2.0}^{3.6} (\sin x + e^x - \log x)\, dx = 15.70393\ 8463, \quad \text{with} \quad h = 0.1.$$

The result is given in Table 1; it is too high by $5.3.10^{-8}$. Weddle's rule, on the other hand, is too high by $2.2.10^{-8}$, and is thus only slightly better than the five-point formula. For comparison, the value obtained by Steffensen's five point formula is also listed.

Example 3. The evaluation of $\int_{2.0}^{4.4} dx/x^2$, $h = 0.1$, results in the value of 0.27272 72723, while the exact value is 0.27272 72727. The same integral, computed in [3] led to 0.27272 72751. The respective errors and estimated errors are again listed in Table 1.

5. Program for Computing Machines. The formulas derived above are clearly well suited for use on modern high-speed computing machinery, such as ORDVAC, SEAC, etc.

For illustrative purposes the programming of formula $I_s = \int_0^{s-1} I_{k,k+1}$ with $I_{k,k+1} = \int_{x_k}^{x_{k+1}} f(x)\,dx$ computed by (2.8) will be discussed briefly. If we are dealing with a machine having a fixed fractional (i.e. decimal, binary) point some preliminary scaling is necessary. Let it be supposed that the machine can handle only numbers of absolute value not exceeding unity.

If, then, $|f| < M$ in an interval containing the total range of integration, we define

$$\bar{f} = f/2M,\ \bar{a}_i = a_i/21{,}600$$

$$\bar{g}_{ki} = \bar{f}_{k+i} + \bar{f}_{k-i+1},\ \bar{I}_{k,k+1} = \sum_{i=1}^{4} \bar{a}_i \bar{g}_{ki} \approx (2Mh)^{-1} I_{k,k+1}.$$

Clearly

$$|\bar{f}| < 2^{-1},\ |\bar{a}_i| < 1,\ |\bar{g}_{ki}| < 1,$$

and since $|I_{k,k+1}| < hM$, also $|\bar{I}_{k,k+1}| < 2^{-1}$.

The build-up of $\bar{I}_{k,k+1}$ is most simply accomplished by first forming the $p_{ki} = \bar{a}_i \bar{g}_{ki}$, then $Q_{ki} = p_{ki} + Q_{k+i+1}$ for $i = 4, 3, 2, 1$, with $Q_{k5} = 0$. Then $Q_{k1} = \bar{I}_{k,k+1}$.

For an efficient accumulation of the $\bar{I}_{k,k+1}$ into $\bar{I}_s$ the "sliding scaling" scheme may be used, as follows: Let

$$\begin{aligned} H_{k+1} &= \bar{I}_{k,k+1}\, 2^{-\alpha_k} + J_k \\ \beta_{k+1} &= 0, 1 \text{ depending on whether } |H_{k+1}| < 2^{-1} \text{ or } \geqq 2^{-1} \qquad (5.1) \\ J_{k+1} &= 2^{-\beta_{k+1}} H_{k+1}, \qquad \alpha_{k+1} = \alpha_k + \beta_{k+1} \end{aligned}$$

for $k = 0, 1, \cdots, s-1$ with $J_0 = 0$, $\alpha_0 = 1$.

It is an easy matter to realize that

$$H_{k+1} = \bar{I}_{k+1}\, 2^{-\alpha_k},\ J_{k+1} = \bar{I}_{k+1}\, 2^{-\alpha_{k+1}}. \qquad (5.2)$$

For $k = 0$, namely, $J_0 = 0$ by definition; thus $\beta_1 = 0$, $\alpha_1 = 1$, and $J_1 = \bar{I}_1 2^{-\alpha_1}$.

If we assume, inductively, that for arbitrary $k < s-1$, $J_k = \bar{I}_k 2^{-\alpha_k}$, then by (5.1) clearly $H_{k+1} = \bar{I}_{k+1} 2^{-\alpha_k}$, whence (5.2) follows immediately.

Further, $\alpha_k = \alpha_0 + \sum_{j=1}^{k} \beta_j$, so that $1 \leqq \alpha_k \leqq k$ for all $k > 0$. Now for $k = 0, 1$ obviously $|J_k| < 2^{-1}$. From $|J_k| < 2^{-1}$, in general, it may be inferred from the equation for H_{k+1} in (5.1) that always $|H_{k+1}| < 1$. Therefore,

$$|J_{k+1}| = 2^{-\beta_{k+1}} |H_{k+1}| < 2^{-1}.$$

After s integrating steps there is obtained $J_s = \bar{I}_s\, 2^{-\alpha_s}$, so that

$$I_s \approx J_s hM 2^{\alpha_s + 1}, \qquad \alpha_s = \alpha_0 + \sum_{j=1}^{s} \beta_j. \qquad (5.3)$$

A rounding error η in J_s is thus translated into a rounding error $hM\eta 2^{\alpha_s+1}$ in I_s.

In order to keep the rounding error R in I_s less than its truncation error it is sufficient, then, to insure $\eta < |R|/hM2^{+\alpha_s+1}$.

The application of the scaling procedure is demonstrated below. Suppose it is desired to compute the integral of example 2 in section 4, again at $h = 0.1$, by means of the seven-point formula (2.8).

TABLE 2

Scaled integration

x	f^{VI}	$\bar{f}$	$I_{k,k+1}$	β_{k+1}	J_{k+1}
2.7		.15099 008461			
2.8		.16613 940405			
2.9		.18300 299537			
3.0		.20177 262281			
3.1		.22265 959740	.21203 030931	0	.10601 515466
3.2		.24589 667980	.23407 199320	0	.22305 11526
3.3		.27174 019788	.25858 989065	0	.35234 609658
3.4	30	.30047 239992	.28585 308311	0	.49527 263813
3.5	33	.33240 406922	.31615 785690	1	.32667 578329
3.6	37	.36787 742780	.34983 045992	0	.41413 339827
3.7	41	.40726 936075			
3.8	45	.45099 499510			
3.9	49	.49951 167080			

In order to estimate the accuracy obtainable it is necessary to compute bounds for the sixth derivative f^{VI}, which is, in this case, approximately e^x.

These approximate values are listed in column 2 of Table 2. They show that

$$R_{7,6} = -\tfrac{1}{100.800}h^7\sum_1^6 f^{VI}(\xi_s) \approx -2 \times 10^{-10}.$$

Further, $|f(x)| < 47.4 \equiv M$ in the interval $2.7 \leqq x \leqq 3.9$.

A preliminary inspection of the J_k reveals them to be sufficiently small to permit setting $\beta_k = 0$ for all k, except for β_5. Thus $\alpha_6 = 2$, and $hM2^{\alpha_6+1} \approx 40$. The rounding error R in I_6 then will not exceed the truncation error if the rounding error η in J_6 is kept below 5×10^{-12}. The use of eleven to twelve decimals throughout the computation is thus indicated.

The values of $\bar{f} = f/2M$ are listed in the third column of Table 2, to eleven decimals. Also,

$$\bar{a}_1 = .56481\ 481481 \qquad \bar{a}_2 = -.07875\ 000000$$

$$\bar{a}_3 = .01550\ 925926 \qquad \bar{a}_4 = -.00157\ 407407.$$

It is seen that $J_6 = .41413\ 339827$, and, consequently,

$$I_6 = 15.70393\ 8462,$$

which is in error by only 1×10^{-9}.

REFERENCES

1. M. Lotkin, A New Integrating Procedure of High Accuracy, J. Math. and Phys. **31**, 29–34, 1952.
2. L. M. Milne-Thomson, The Calculus of Finite Differences, London 1951.
3. G. Blanch and I. Rhodes, Seven-Point Lagrangian Integration Formulas, J. Math. and Phys. **22**, 204–207, 1943.
4. P. M. Hummel and C. L. Seebeck, A Generalization of Taylor's Expansion, Am. Math. Monthly, **56**, 243–247, 1949.
5. A. A. Bennett, W. E. Milne, H. Bateman, Numerical Integration of Differential Equations, National Research Council, Bulletin No. 92, 1933.
6. W. E. Milne, A Note on the Numerical Integration of Differential Equations, NBS Journal of Research, **43**, 537–542, 1949.

Ballistic Research Laboratories
Aberdeen Proving Ground

(Received September 12, 1952)

THE VECTOR METHOD OF SOLVING SIMULTANEOUS LINEAR EQUATIONS

By Everett W. Purcell

This paper is written to present a new method of solving simultaneous linear equations which has some advantages over other methods when equipment such as the IBM Card Programmed Calculator is available.

It will be convenient for us to write our equations as follows:

$$\text{(I)} \qquad \begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n + a_{1,n+1}t &= 0 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n + a_{2,n+1}t &= 0 \\ &\cdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n + a_{n,n+1}t &= 0 \end{aligned}$$

where we have introduced the parameter t to make the equations homogeneous. (Later on we shall arrange to have $t = 1$.)

Let A denote the $n \times (n+1)$ matrix of coefficients and let V_i denote the ith row of A.

$$\text{Then } A = \begin{vmatrix} V_1 \\ V_2 \\ \cdot \\ \cdot \\ \cdot \\ V_n \end{vmatrix}$$

Also let V_i^j denote the ith vector which is orthogonal to the first j vectors of matrix A; i.e., the coordinates of V_i^j will satisfy the first j of our equations I and it is the ith set of such coordinates. It is now clear that any linear combination of the V_i^j will also be a solution of the first j equations and linear combinations exist which will satisfy other equations in I.

We shall now outline one method of finding linear combinations and ultimately a solution for our equations which is simple and convenient. First we shall define V_i^0 to be an $n+1$ dimensional row vector with all coordinates equal to zero except the ith which is 1.

First we shall find C_i^1 such that $V_i^1 = C_i^1 V_1^0 + V_{1+i}^0$ will be orthogonal to V_1; i.e.,

$$V_1 \cdot V_i^1 = 0 = C_i^1 V_1 \cdot V_1^0 + V_1 \cdot V_{1+i}^0$$

therefore

$$\text{(1)} \qquad C_i^1 = -(V_1 \cdot V_{1+i}^0 / V_1 \cdot V_1^0) = -(a_{1,1+i}/a_{11}) \qquad i = 1, 2, \cdots, n$$

Next we define $V_i^2 = C_i^2 V_1^1 + V_{i+1}^1$ and similarly determine C_i^2 such that V_i^2

are orthogonal to V_2. This gives

$$(2)\quad C_i^2 = -(V_2 \cdot V_{i+1}^1 / V_2 \cdot V_1^1) = -(a_{21} C_{1+i}^1 + a_{2,2+i})/(a_{21} C_1^1 + a_{22})$$

$$i = 1, 2, \cdots, n-1$$

We now have $n-1$ vectors V_i^2 all of which are orthogonal to V_1 and V_2. We proceed in this manner at each stage defining $V_i^{j+1} = C_i^{j+1} V_1^j + V_{i+1}^j$ and computing C_i^j from

$$(3)\qquad C_i^j = -(V_j \cdot V_{1+i}^{j-1} / V_j \cdot V_1^{j-1})$$

We ultimately end up with one vector which is orthogonal to all the V_i. Thus we have a solution to our system of equations. The above procedure is such that we end up with $t = 1$, but other procedures exist for which we will have to normalize our last vector making the t coordinate equal to one before we have the desired solution.

It can be shown that the number of multiplications needed to solve an $n \times n$ system by the vector method is given by

$$\tfrac{1}{6}(n-1)(n)(2n+5).$$

This is the same as that needed when Crout's Method is used. It is easy to see that the number of additions and divisions is the same for both methods.

Although the amount of arithmetic is the same for both methods, the number of storage cells required is less for the Vector Method than for the Crout Method. It can be shown that the number of storage cells given by the greatest integer in

$$n^2/4 + n + 2$$

is sufficient to stove computed results in the programming of a solution on the CPC when using the Vector Method.

We shall now show a method of evaluating determinants. First let us observe that if we let

$$C_i^j = +N_i^j / D_j,$$

we can now write

$$N_i^2 = -\begin{vmatrix} 1 & -C_{i+1}^1 \\ a_{21} & a_{2,2+i} \end{vmatrix} \qquad D_2 = \begin{vmatrix} 1 & -C_1^1 \\ a_{21} & a_{22} \end{vmatrix}$$

$$V_i^2 = \left\{ \begin{vmatrix} -C_1^1 & -C_{i+1}^1 \\ 1 & -C_i^2 \end{vmatrix}, -\begin{vmatrix} 1 & -C_{i+1}^1 \\ 0 & -C_i^2 \end{vmatrix}, 0, \cdots, 1, 0, \cdots 0 \right\}$$

where the $(i+2)$nd coordinate is 1. Similarly

$$N_i^3 = -\begin{vmatrix} 1 & -C_1^1 & -C_{i+2}^1 \\ 0 & 1 & -C_{i+1}^2 \\ a_{31} & a_{32} & a_{3,3+i} \end{vmatrix} \qquad D_3 = \begin{vmatrix} 1 & -C_1^1 & -C_2^1 \\ 0 & 1 & -C_1^2 \\ a_{31} & a_{32} & a_{33} \end{vmatrix}$$

$$V_i^3 = \left\{ -\begin{vmatrix} -C_1^1 & -C_2^1 & -C_{2+i}^1 \\ 1 & -C_1^2 & -C_{1+i}^2 \\ 0 & 1 & -C_i^3 \end{vmatrix}, \begin{vmatrix} 1 & -C_2^1 & -C_{2+i}^1 \\ 0 & -C_1^2 & -C_{1+i}^2 \\ 0 & 1 & -C_i^3 \end{vmatrix}, \right.$$

$$\left. -\begin{vmatrix} 1 & -C_1^1 & -C_{2+i}^1 \\ 0 & 1 & -C_{1+i}^2 \\ 0 & 0 & -C_i^3 \end{vmatrix} 0, \cdots, 1 \right\}$$

where the $(i+3)$rd coordinate is 1. We can now deduce the rule for the formation of the scalar multipliers C_i^j and the vectors V_i^j.

It can be proved by induction that

$$D_j = \begin{vmatrix} 1 & -C_1^1 & -C_2^1 & \cdots & -C_{j-1}^1 \\ 0 & 1 & -C & \cdots & -C_{j-2}^2 \\ 0 & 0 & 1 & & -C_{j-2}^3 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{j1} & a_{j2} & a_{j3} & \cdots & a_{jj} \end{vmatrix} = V_j \cdot V_1^{j-1}$$

and that N_i^j can be gotten from D_j by adding i to the subscripts in the last column. Also

$$V_i^j = \{ |a_{j+1,1}|, |a_{j+1,2}|, \cdots, |a_{j+1,j}|, 0, \cdots 1, \cdots \}$$

where the $(j+i)$th coordinate is 1. By $|a_{j+1,i}|$ we mean the minor of $a_{j+1,i}$ in the determinant which we may denote by $D_{j+1} = V_{j+1} \cdot V_1^j$. It is now easy to see that

$$D = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix} = D_1 \begin{vmatrix} 1 & -C_1^1 & -C_2^1 & \cdots & -C_{n-1}^1 \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} \end{vmatrix}$$

$$= D_1 \cdot D_2 \cdot \begin{vmatrix} 1 & -C_1^1 & -C_2^1 & \cdots & -C_{n-1}^1 \\ 0 & 1 & -C_1^2 & \cdots & -C_{n-2}^2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & \cdots & a_{nn} \end{vmatrix} = D_1 \cdot D_2 \cdots \cdot D_n$$

THEOREM: Let D be a determinant of order n; then

$$D = \prod_{i=1}^{n} D_i = \prod_{i=1}^{n} V_i \cdot V_1^{i-1}$$

Example

$$\begin{aligned} u + x + y + z - 1 &= 0 \\ 2u + x + 2y + z - 1 &= 0 \\ 2u + 2x - y + 2z - 5 &= 0 \\ u + 2x + 3y + 3z - 3 &= 0 \end{aligned}$$

First we get

$V_1^1 = (-1, 1, 0, 0, 0)$, $V_2^1 = (-1, 0, 1, 0, 0)$, $V_3^1 = (-1, 0, 0, 1, 0)$, $V_4^1 = (1, 0, 0, 0, 1)$

These four vectors are orthogonal to $V_1 = (1, 1, 1, 1, -1)$. Next, using $V_2 = (2, 1, 2, 1, -1)$, we get

$$C_1^2 = 0; C_2^2 = -1; C_3^2 = 1$$

giving

$$V_1^2 = (-1, 0, 1, 0, 0), V_2^2 = (0, -1, 0, 1, 0), V_3^2 = (0, 1, 0, 0, 1)$$

Then $C_1^3 = 0; C_2^3 = -1$

$$V_1^3 = (0, -1, 0, 1, 0), V_2^3 = (1, 1, -1, 0, 1)$$

Now $C_1^4 = 3$, giving $V_1^4 = (1, -2, -1, 3, 1)$

$$\therefore u - 1, x = -1, y = -1, z = 3$$

REFERENCES

1. Prescott D. Crout, *A Short Method for Evaluating Determinants and Solving Systems of Linear Equations with Real or Complex Coefficients, ...* Trans. American Inst. El. Eng. **60**, 1941.
2. W. E. Milne, *Numerical Calculus ...* pp. 15–23.
3. C. C. MacDuffee, *Vectors and Matrices ...* Carus Mathematical Monograph No. 7. pp 1–30
4. G. Birkhoff and S. MacLane, *A Survey of Modern Algebra.* ch 7

438 North G Street
Oxnard, California

(Received December 19, 1952)

ON THE SOLUTION OF THE DIFFERENTIAL EQUATION OCCURRING IN THE PROBLEM OF HEAT CONVECTION IN LAMINAR FLOW THROUGH A TUBE*

By Milton Abramowitz

1. Introduction. The problem of forced heat convection in laminar flow through a circular tube or duct assuming the velocity distribution of the fluid to be the Poiseuille flow requires the solution of the ordinary differential equation [1, 2].

$$d^2y/dr^2 + (1/r)\, dy/dr + \beta^2(1 - r^2)y = 0 \tag{1.1}$$

subject to the end conditions

$$y(0) = 1 \tag{1.2}$$

$$y(1) = 0. \tag{1.3}$$

The condition (1.2) arises from the physical condition that the temperature shall be constant at the mouth of the tube, while (1.3) results from the fact that there shall be no heat transfer at the surface of the tube.

A particular integral of (1.1) satisfying (1.2) is

$$y(r, \beta) = \sum_{n=0}^{\infty} B_{2n} r^{2n} \tag{1.4}$$

where

$$\begin{aligned} &B_0 = 1,\ 4B_2 + \beta^2 = 0 \\ &4n^2 B_{2n} + \beta^2[B_{2n-2} - B_{2n-4}] = 0 \end{aligned} \tag{1.5}$$

In order to satisfy (1.3) the parameter β must take on an infinite number of discrete values β_s. One of the procedures employed in determining the proper values of β is to tabulate the function $y(r, \beta)$ with the aid of (1.4) for a sequence of values of β in the neighborhood of β_s and then determine β_s from that value of β for which $y(1, \beta_s) = 0$. This technique has certain disadvantages computationally. It is evident from (1.5) that the individual terms of (1.4) increase as β_s^2 increases. The difficulty in determining β_s from the condition $y(1, \beta_s) = 0$ is that this equation is satisfied to a high degree of accuracy even when β_s is not known very precisely. Thus, if we take $\beta_0 = 2.7$, $y(1, 2.7) = .00219$. Therefore, the determination of precise values of β_s requires calculations of $y(1, \beta_s)$ to a large number of decimal places. It is the purpose of the present article to obtain a solution of (1.1) in terms of Bessel functions which shall be more convenient for the determination of the values of β_s.

2. Solution of (1.1), in terms of Bessel functions. In order to determine a solution of the differential equation (1.1) as a series of Bessel functions let us formally assume the possibility of an expression in the form

$$y(r, \beta) = \sum_{n=0}^{\infty} A_n(\beta)\Omega_n(\beta r) \tag{2.1}$$

where

* The preparation of this paper was sponsored in part by the Flight Research Laboratory, Wright Air Development Center, USAF.

$$\Omega_n(\beta r) = (\beta r)^n J_n(\beta r) \tag{2.2}$$

The functions $\Omega_n(\beta r)$ satisfy the following recurrence relations [3] where the differentiation is with respect to the argument βr.

$$(1/r)\Omega_n' = \beta\Omega_{n-1} \tag{2.3}$$

$$\beta^2 r^2 \Omega_n = (2n + 2)\Omega_{n+1} - \Omega_{n+2} \tag{2.4}$$

$$\Omega_n'' = (2n - 1)\Omega_{n-1} - \Omega_n \tag{2.5}$$

The Ω_n were chosen to represent the solution of (1.1) since they are even functions of r and the recurrence relation (2.4) has the property that it expresses $r^2\Omega_n$ as a linear combination of functions of different orders. Thus, substituting (2.1) into (1.1) we obtain

$$\sum_{n=0}^{\infty} A_n(\beta)\{\beta^2\Omega_n''(\beta r) + (\beta/r)\Omega_n'(\beta r) + \beta^2\Omega_n(\beta r) - \beta^2 r^2\Omega_n(\beta r)\} = 0 \tag{2.6}$$

With the aid of (2.3)–(2.5) this last relation becomes

$$\sum_{n=0}^{\infty} A_n(\beta)\{(2n - 1)\beta^2\Omega_{n-1} - \beta^2\Omega_n + \beta^2\Omega_{n-1} + \beta^2\Omega_n - (2n + 2)\Omega_{n+1} + \Omega_{n+2}\} = 0 \tag{2.7}$$

which becomes

$$\sum_{n=0}^{\infty} A_n(\beta)\{2n\beta^2\Omega_{n-1} - (2n + 2)\Omega_{n+1} + \Omega_{n+2}\} = 0 \tag{2.8}$$

Equating the coefficients of Ω_n to zero we get the following recurrence relations for the coefficients $A_n(\beta)$.

$$A_1 = 0,\ 2\beta^2 A_2 - A_0 = 0,\ 6\beta^2 A_3 + A_0 = 0,\ 4\beta^2 A_4 - 3A_2 = 0 \tag{2.9}$$

$$(2n + 2)\beta^2 A_{n+1} - 2nA_{n-1} + A_{n-2} = 0 \qquad \text{for } n \geqq 4 \tag{2.10}$$

In order to satisfy (1.2) we must choose $A_0 = 1$ and we then have for the first five coefficients

$$A_0 = 1,\ A_1 = 0,\ A_2 = 1/(2\beta^2),\ A_3 = -1/(6\beta^2),\ A_4 = 3/(8\beta^4) \tag{2.11}$$

3. Determination of the values of β_s. In order to determine the values of β_s we must determine the zeros of the function

$$y(r, \beta) = \sum_{n=0}^{\infty} A_n(\beta)(\beta r)^n J_n(\beta r) \tag{3.1}$$

for $r = 1$. This was done by evaluating (3.1) for a series of values in the neighborhood of each β_s and then evaluating the zero by inverse interpolation. For example for β_4 we find

β	$y(1, \beta)$
18.65	+.0052576
18.66	+.0026192
18.67	−.0000339

By linear inverse interpolation we get $\beta_4 = 18.66987$. Below are the first five values that were obtained in this fashion.

s	β_s
0	2.7043644
1	6.679032
2	10.67338
3	14.67108
4	18.66987

In carrying out the computations the trial values of β were so chosen that the values $J_n(\beta r)$ for $r = 1$ could be obtained from existing tables of the Bessel functions J_n without recourse to interpolation. The calculations were carried to at least two more decimal places than given in the final computed values of β_s. The number of terms carried was such that the truncating error in the series (2.1) was less than the accuracy required in the values $y(1, \beta)$ for evaluating the β_s by inverse interpolation. A sufficient number of trial values were taken so that the calculations could be checked by differencing. The choice of the trial values is worth mentioning. An examination of the values of β_0 and β_1 shows that they appear to tend to $4s + 2 + \frac{2}{3}$ and this was used in determining the higher characteristic values.

In addition to the values of β_s the coefficients $\beta_s^n A_n(\beta_s)$ necessary for the determination of $y(r, \beta_s)$ to six decimal places have also been given. The quantities $\beta_s^n A_n(\beta_s)$ have been given rather than the $A_n(\beta_s)$ themselves because they are more convenient for computation due to their slower rate of change. The coefficients themselves are correct to probably no more than seven significant figures.

The problem has been discussed by Lauwerier [4] who obtained an asymptotic expansion for the characteristic values. His result in the notation used here is

$$\tfrac{1}{2}\beta_s = v + 0.031580\, v^{-4/3} - 0.000904\, v^{-8/3}$$

where $v = 2s + 1 + \frac{1}{3}$. Comparing the results with that obtained by Lauwerier [4] we have

s	$\frac{1}{2}\beta_s$	*Asymptotic Formula*
0	1.3521822	1.354
1	3.339516	3.3396
2	5.33669	5.33671
3	7.33554	7.33554

The asymptotic formula will thus be adequate for the determination of the higher characteristic values.

REFERENCES

1. Jakob, M., Heat Transfer, vol. 1. New York: John Wiley, 1949.
2. Boelter, L. M. K. et al., Heat Transfer Notes. University of California Press, Berkeley, Calif., 1948.
3. Watson, G. N., Theory of Bessel Functions, second edition. Cambridge University Press, 1944, p. 45.
4. Lauwerier, H. A., The Use of Confluent Hypergeometric Functions in Mathematical Physics and the Solution of the Eigenvalue Problem. Applied Scientific Research, 1951, A, 2, pp. 184–204.

National Bureau of Standards

(Received November 26, 1952)

TABLE 1

$\beta_0 = 2.7043644$; $\beta_1 = 6.679032$; $\beta_2 = 10.67338$; $\beta_3 = 14.67108$; $\beta_4 = 18.66987$

n	$\beta_0{}^nA_n$	$\beta_1{}^nA_n$	$\beta_2{}^nA_n$	$\beta_3{}^nA_n$	$\beta_4{}^nA_n$
0	1.0000000	1.000000	1.000000	1.000000	1.000000
1	.0000000	.000000	.000000	.000000	.000000
2	.5000000	.500000	.500000	.500000	.500000
3	−.4507274	−1.113172	−1.778897	−2.445180	−3.111645
4	.3750000	.375000	.375000	.375000	.375000
5	−.4958001	−1.224489	−1.956786	−2.689698	−3.422810
6	.4140776	.932076	1.894737	3.301953	5.153667
7	−.4974099	−1.228465	−1.963140	−2.698431	−3.433923
8	.4461194	1.326717	2.963240	5.355507	8.503422
9	−.5043542	−1.437822	−2.868527	−5.089895	−8.397837
10	.4687663	1.604293	3.714583	6.799401	10.858624
11	−.5133433	−1.709892	−4.045378	−8.198589	−14.850660
12		1.870738	4.680737	9.344211	16.486510
13		−1.990482	−5.259085	−11.404642	−21.505574
14		2.144086	5.888465	12.972559	25.211041
15		−2.274274	−6.573789	−15.213988	−30.331903
16		2.426377	7.274568	17.390475	36.182422
17		−2.561859	−8.035620	−19.916738	−42.391405
18			8.819440	22.624494	49.902640
19			−9.655962	−25.582621	−57.937150
20			10.522649	28.798271	67.193558
21			−11.437422	−32.267395	−77.360995
22			12.386659	36.019365	88.722920
23			−13.381713	−40.049285	−101.269104
24			14.413797	44.381006	115.116126
25				−49.016174	−130.347248
26				53.973396	147.047833
27				−59.258487	−165.319629
28				64.887207	185.252663
29				−70.867643	−206.952813
30				77.214067	230.519174
31				−83.935871	−256.061483
32					283.687046
33					−313.510575
34					345.646759
35					−380.215990
36					417.339928
37					−457.145018

EVALUATION OF THE INTEGRAL $\int_0^\infty e^{-u^2-x/u}\,du$

By Milton Abramowitz

1. Introduction.[1] The function $f_m(x)$ defined by the integral

$$f_m(x) = \int_0^\infty u^m e^{-u^2-x/u}\,du, \qquad m = 0, 1, 2, \cdots, \tag{1.1}$$

appears in various physical problems where particles having Maxwellian distributions of velocity are involved. Thus the fraction of a parallel beam of thermal neutrons which is transmitted to an absorber (following an inverse velocity law) as a function of thickness is given by $f_3(x)$, with x real and greater than zero. For this and other thermal neutron absorption problems, C. T. Zahn [1] has obtained power series for $f_1(x)$, $f_2(x)$, and $f_3(x)$, and has tabulated these three functions for $0 \leqq x \leqq 1$. Laporte [2] has found the first three terms of an asymptotic expansion for $f_m(x)$, but the coefficient of the third term is in error, which is noted by H. C. Torrey [3]. The latter finds the shape of a radiofrequency spectral line to be representable (except for a simple multiplicative function) by $f_3(x)$ with x a pure imaginary number. He gives a recurrence formula for the coefficients of the terms in the asymptotic expansion of $f_m(x)$. A misprint in the recurrence formula is noted by U. E. Kruse and N. F. Ramsey [4], who present a formula for the case $m = 3$. Their formula in turn contains a misprint, which we correct in Section (4) of the present paper. They tabulate $f_3(-iy)$ for $0 \leqq y \leqq 20$.

The modulation of the momenta of the molecules of a gas, on being reflected by a wall vibrating sinusoidally with time, varies with distance from the wall. In the simple case of no collisions between molecules, the variation of the modulation with distance from the wall is given by $f_1(x)$, with x a pure imaginary number.

It seems desirable to re-examine the algorithms for the computation of $f_m(x)$. The function $f_m(x)$ satisfies the following differential equation and recurrence relations,

$$xf_m''' - (m-1)f_m'' + 2f_m = 0, \tag{1.2}$$

$$f_m' = -f_{m-1} \tag{1.3}$$

$$2f_m = (m-1)f_{m-2} + xf_{m-3}\,. \tag{1.4}$$

However, for simplicity we shall restrict ourselves to the case $m = 0$, since the methods described may be employed for the general case in exactly the same manner.

[1] The author is indebted to R. K. Cook and M. Greenspan of the Sound Section of the National Bureau of Standards who suggested the problem and wrote the introductory section.

2. Statement of problem: Series Expansion. We are concerned with the problem of evaluating the integral for $f_0(x)$,

$$f_0(x) = \int_0^\infty e^{-u^2 - x/u}\, du, \tag{2.1}$$

the subscript being omitted for convenience in what follows.

For applications it is desirable to have expressions which will be convenient for all values of x. We shall obtain two representations of the function, the first suitable for small values of x and the second for large values of x. We note first that $f(0) = \sqrt{\pi}/2$ and further that $f(x)$ satisfies the ordinary linear differential equation of the third order.

$$xf''' + f'' + 2f = 0 \tag{2.2}$$

subject to the condition

$$f(0) = \sqrt{\pi}/2. \tag{2.3}$$

Examination of (2.2) shows that $f(x)$ must have the form $a(x)\cdot\log x + b(x)$ where $a(x)$ and $b(x)$ are power series. An attempt to solve (2.2) by series will not be convenient because of the difficulty of determining the integration constants. However the desired expression can be obtained readily with the aid of the Laplace transform. The Laplace transform of a function $f(x)$ is defined as

$$\mathfrak{L}\{f(x)\} = \int_0^\infty e^{-px} f(x)\, dx = G(p) \tag{2.4}$$

Thus operating on (2.1) we then obtain by inverting the order of integration which is permissible because of the absolute convergence of (2.1),

$$\mathfrak{L}\{f(x)\} = G(p) = \frac{1}{p}\int_0^\infty \frac{u}{u + 1/p}\, e^{-u^2}\, du \tag{2.5}$$

Since $u/(u + 1/p) = 1 - (1/p)(1/[u + 1/p])$ we can rewrite (2.5) in the form

$$G(p) = \frac{1}{p}\int_0^\infty e^{-u^2}\, du - \frac{1}{p^2}\int_0^\infty \frac{e^{-u^2}}{u + 1/p}\, du \tag{2.6}$$

The series expansion of the second integral in (2.6) (which can also be obtained by operational methods) is known [5] and we have

$$\begin{aligned}\int_0^\infty \frac{e^{-u^2}}{u + 1/p}\, du &= \tfrac{1}{2}\sum_{n=0}^\infty \frac{(-1)^n \psi(n+1)}{n!}\left(\frac{1}{p}\right)^{2n} \\ &\quad + \sqrt{\pi}\sum_{n=0}^\infty \frac{(-2)^n}{1\cdot 3\cdot 5\cdots(2n+1)}\left(\frac{1}{p}\right)^{2n+1} + e^{-1/p^2}\log p\end{aligned} \tag{2.7}$$

where $\psi(n+1) = 1 + 1/2 + 1/3 + \cdots + 1/n - \gamma, \psi(1) = -\gamma$ and $\gamma = .577\cdots$ is Euler's constant. Expanding the exponential in the last term and making use

of the fact that $f(0) = \sqrt{\pi}/2$ we get

$$G(p) = \frac{1}{p}\frac{\sqrt{\pi}}{2} - \sum_{n=0}^{\infty} \frac{(-1)^n\{\frac{1}{2}\psi(n+1) + \log p\}}{n!\, p^{2n+2}} - \sqrt{\pi}\sum_{n=0}^{\infty} \frac{(-2)^n}{1\cdot 3\cdot 5\cdots(2n+1)p^{2n+2}} \tag{2.8}$$

Now making use of the transforms

$$\mathfrak{L}\left\{\frac{x^k}{k!}\right\} = \frac{1}{p^{k+1}}, \qquad \mathfrak{L}\left\{\frac{x^k}{k!}[\psi(k+1) - \log x]\right\} = \frac{\log p}{p^{k+1}}$$

we get immediately as the inverse of $G(p)$,

$$f(x) = \frac{\sqrt{\pi}}{2} - \sum_{n=0}^{\infty} \frac{(-1)^n\{\frac{1}{2}\psi(n+1) + \psi(2n+2) - \log x\}}{n!\,(2n+1)!}\, x^{2n+1} - \sqrt{\pi}\sum_{n=0}^{\infty} \frac{(-2)^n x^{2n+2}}{1\cdot 3\cdot 5\cdots(2n+1)[(2n+2)!]} \tag{2.9}$$

3. Asymptotic Expansion. In order to determine an expression for $f(x)$ suitable for large values of x we make use of the differential equation (2.2). However, it will be necessary to make a change of variable in (2.2) which we can justify by approximating the integral (2.1) by the method of saddle points. Thus, let us write

$$u = tv,\ t = (x/2)^{1/3} \tag{3.1}$$

and (2.1) becomes

$$f(x) = t\int_0^{\infty} e^{-t^3(v^2+2/v)}\, dv \tag{3.2}$$

It can now be shown (Laporte [2]) by the saddle point method that if x is large then this integral may be approximated by

$$f(x) \sim te^{-3t^3}\int_0^{\infty} e^{-3t^3(v-1)^2}\, dv \tag{3.3}$$

Since the principal contribution to the integral comes in the neighborhood of $v = 1$ we have

$$f(x) \sim \frac{e^{-3t^3}}{\sqrt{3}}\int_{-\infty}^{\infty} e^{-s^2}\, ds = \sqrt{\frac{\pi}{3}}\, e^{-3t^3} = \sqrt{\frac{\pi}{3}}\, e^{-3(x/2)^{3/3}} \tag{3.4}$$

as the asymptotic behavior of $f(x)$. This expression suggests the following substitution in (2.2). Let

$$z = 3(x/2)^{2/3} \tag{3.5}$$

and we get ultimately

$$\frac{d^2f}{dz^2} + \frac{1}{4z^2}\frac{df}{dz} + f = 0 \tag{3.6}$$

Now, if $z \to \infty$, (3.6) reduces to $f''' + f = 0$ for which the desired particular solution is e^{-z}. We therefore assume that $f(z)$ can be expressed in the form

$$f(z) = \sqrt{\pi/3}e^{-z}U(z) \tag{3.7}$$

the constant being fixed by (3.4). The differential equation for $U(z)$ is

$$U''' - 3U'' + \left[3 + \frac{1}{4z^2}\right]U' - \frac{1}{4z^2}U = 0 \tag{3.8}$$

If we now assume an expansion of the form

$$U(z) = \sum_{n=0}^{\infty} a_n z^{-n}, \qquad a_0 = 1 \tag{3.9}$$

and substitute in (3.8) we obtain the following recurrence relations for the coefficients a_n,

$$a_1 = -1/12,\ a_2 = 25/288 \tag{3.10}$$

$$(3n + 6)a_{n+2} + [3(n + 1)(n + 2) + 1/4]a_{n+1} + [n(n + 1)(n + 2) + n/4]a_n = 0$$

The expression for $f(x)$ up to the term in z^{-3} is

$$f(x) = \sqrt{\frac{\pi}{3}}\, e^{-3(x/2)^{2/3}}\left\{1 - \frac{1}{36}\left(\frac{2}{x}\right)^{2/3} + \frac{25}{2592}\left(\frac{2}{x}\right)^{4/3} - \frac{1225}{279936}\left(\frac{2}{x}\right)^{2} + \cdots\right\}. \tag{3.11}$$

It should be mentioned that the result given in (3.11) can be obtained with the aid of the Laplace transform operating on (3.6). The operational form of (3.6) then can be reduced to the hypergeometric equation from which we can ultimately derive both (3.11) and (2.9) by choosing the appropriate particular solution. The methods employed here are preferable because they are simpler.

4. The General Case. Following Laporte [2] it can be shown that an asymptotic expansion of $f_m(x)$ is

$$f_m(x) \sim \sqrt{\frac{\pi}{3}}\, 3^{-m/2} z^{m/2} e^{-z}\left(a_0 + \frac{a_1}{z} + \frac{a_2}{z^2} + \cdots + \frac{a_\mu}{z^\mu} + \cdots\right) \tag{4.1}$$

where $z = 3(x/2)^{2/3}$, and in which the a's are functions of m. We present certain properties of the a's mainly for the purpose of correcting the published errors referred to in Section 1.

The values of a_0 and a_1 are

$$a_0 = 1,\ a_1 = 1/12(3m^2 + 3m - 1) \tag{4.2}$$

The general recurrence relation is

$$12(\mu + 2)a_{\mu+2} = -(12\mu^2 + 36\mu - 3m^2 - 3m + 25)a_{\mu+1} + \tfrac{1}{2}(m - 2\mu)(2\mu + 3 - m)(2\mu + 3 + 2m)a_\mu,\ \mu = 0, 1, 2, 3, \cdots \tag{4.3}$$

From (4.3) and (4.2) we find

$$a_2 = \tfrac{1}{288}(9m^4 + 6m^3 - 51m^2 - 24m + 25) \tag{4.4}$$

For the case $m = 3$, (4.3) reduces to

$$12\,pa_p = (-12p^2 + 12p + 35)a_{p-1} - \tfrac{1}{2}(7 - 2p)(5 + 2p)(4 - 2p)a_{p-2}, \qquad p = 2, 3, 4, \cdots, \tag{4.5}$$

in which we have used the running index p so as to facilitate comparison with Kruse and Ramsey's result [4].

REFERENCES

1. Zahn, C. T., Absorption coefficients for thermal neutrons. Physical Review **52**, 67 (1937).
2. Laporte, O., Absorption coefficients for thermal neutrons. Physical Review **52**, 72 (1937).
3. Torrey, H. C., Notes on Intensities of Radiofrequency Spectra. Physical Review **59**, 293 (1941).
4. Kruse, U. E. and Ramsey, N. F., The integral $\int_0^\infty y^3 \exp(-y^3 + ix/y)\,dy$, Jour. of Math. and Physics **30**, 40 (1951).
5. Goodwin, E. T. and Staton, J., Table of $\int_0^\infty e^{-u^2}/(u + x)\,du$. Quarterly Journal Mech. Appl. Math. **1**, 319 (1948).

National Bureau of Standards

(Received September 18, 1952)

NOTE ON THE ZEROS OF $(dP^1_{m_i}(x)/dx)\,|_{x=x_0}$

By K. M. Siegel, J. W. Crispin, R. E. Kleinman and H. E. Hunter

In an earlier article the authors [1] presented a method for the determination of the positive elements of the set

$$\{n_i \mid P^1_{n_i}(x_0) = 0\}, \tag{1}$$

the zeros of the Associated Legendre Function of the first kind of order 1 and non-integral degree, and the evaluation of the associated integrals

$$\int_{x_0}^{1} [P_{n_i}(x)]^2\, dx, \qquad (x = \cos\theta). \tag{2}$$

This note presents an extension of the process to the determination of the positive elements of the set

$$\{m_i \mid dP^1_{m_i}(x_0)/dx = 0\}, \tag{3}$$

and the evaluation of the integrals

$$\int_{x_0}^{1} [P^1_{m_i}(x)]^2\, dx, \qquad (x = \cos\theta). \tag{4}$$

Prior to the discussion of the method employed for finding the m_i, let us first examine a variation on MacDonald's formula which will permit a quick approximation to the first m_i for a given θ_0. By differentiating, with respect to μ, equation (14) of Reference 2 one obtains

$$\begin{aligned} dP_n^{-1}(-\mu)/d\mu &= [\cos(n-1)\pi](dP_n^{-1}(\mu)/d\mu) \\ &\quad - (2/\pi)[\sin(n-1)\pi](dQ_n^{-1}(\mu)/d\mu), \end{aligned} \tag{5}$$

where $-\mu = \cos\theta$ and the Q is the Associated Legendre Function of the second kind.

If we assume that $n = m_1$, using (3) and recursion formulas [p. 61, ref. 3], equation (5) reduces to

$$\begin{aligned} &\tan(n-1+k)\pi \\ &\quad = \pi[n(n+1)\tan^2\tfrac{1}{2}(\pi-\theta_0)][n+2-(n+1)\cos(\pi-\theta_0)]/ \\ &\qquad [n-(n+1)\cos(\pi-\theta_0)], \ (k = 0, -1, \cdots). \end{aligned} \tag{6}$$

For values of θ_0 less than and nearly equal to π, if, for a given n, k can be chosen so that $(n-1+k)\pi$ is small, then (6) reduces to the approximation formula.

$$n - 1 + k = -n(n+1)(\pi-\theta_0)^2/4. \tag{7}$$

The application of this formula to the case of $x_0 = \cos 165°$ results in the approximation $m_1 = .967$, which is determined by taking $k = 0$.

The method employed in finding the elements of the set (3) is very similar to the method discussed in reference 1 relative to the elements of the set (1). We assume that m_i is of the form $j + z_j$ or $j + \frac{1}{2} + z_j$ ($j = 0, 1, 2, \cdots$) and use the same Taylor expansions employed in reference 1. With the boundary condition stated in (3) rather than the one in (1) we obtain the following quadratic in z_j

$$\{\tfrac{1}{2}y(y+1)[\partial^2 P_y(x)/\partial y^2] + (2y+1)[\partial P_y(x)/\partial y] - \tfrac{1}{2}x[\partial^3 P_y(x)/\partial x \partial y^2]\}z_j^2$$

$$+ \{y(y+1)[\partial P_y(x)/\partial y] + (2y+1)[P_y(x)] - x(\partial^3 P_y(x)/\partial x \partial y]\}z_j \tag{8}$$

$$+ \{y(y+1)[P_y(x)] - x[\partial P_y(x)/\partial x]\} = 0$$

where the coefficients are evaluated at $x = x_0$ and $y = j$ if the m_1 is assumed to be of the form $j + z_j$ or $y = j + \frac{1}{2}$ if the m_i is assumed to be of the form $j + \frac{1}{2} + z_j$.

Thus equation (8) provides a method for the determination of the z_j and thus a method of obtaining two approximations of each m_i. This is completely analogous to the process employed in determining the elements of the set (1), [ref. 1], and it is worth noting that the work involved in finding the n_i's provides the computations required to evaluate the coefficients in equation (8). The final choice between the two approximations, is based, as in reference 1, upon the smaller value of $|z_j|$.

In reference 1 it was shown that certain bounds and checks could be readily established for the n_i's. Similar checks are available in this case and are obtained from the relation

$$(1 - x_0^2)^{3/2}[dP_n^1(x)/dx]_{x=x_0} = n[(n+1-nx_0^2)P_n(x_0) - x_0 P_{n-1}(x_0)], \tag{9}$$

which can be obtained from the recursion formulas appearing on p. 62 of reference 3. Using the values of $P_n(x_0)$ and $P_{n-1}(x_0)$ determined for $n = \frac{1}{2}K$ (K = an integer) by the methods presented in reference 1 one is able to establish intervals of width $\frac{1}{2}$ which contain the m_i's.

To evaluate the integrals (4) we begin with

$$\int_{x_0}^{1} [P_{m_i}^1(x)]^2\, dx = \left\{m_i(m_i+1)(x_0^2-1)P_{m_i}(x_0)\right\} \cdot \left\{\frac{P_y(x)}{x} + \frac{y(y+1)}{x(2y+1)} \cdot \frac{\partial P_y(x)}{\partial y} - \frac{1}{2y+1}\frac{\partial^2 P_y(x)}{\partial x \partial y}\right\}_{\substack{x=x_0 \\ y=m_i}} \tag{10}$$

[p. 20, ref. 4]

where the m_i's are determined by the process discussed above and thus are of the form $j + z_j$ or $j + \frac{1}{2} + z_j$. For simplicity the integrals are evaluated by both methods in each case; however the final choice is based upon the same criterion used in connection with the m_i's themselves.

The determination of the computational forms required for the evaluation of these integrals is accomplished by continuing the procedure of approximation

outlined above and in reference 1. The forms evolved for this purpose are

$$P_{m_i}(x_0) = P_{N+z_j}(x_0) = \left[\sum_{s=0}^{M+1} \frac{1}{s!} \frac{\partial^s P_y(x)}{\partial y^s} z_j^s\right]_{\substack{x=x_0 \\ y=N}}$$

$$\left.\frac{\partial P_y(x)}{\partial y}\right|_{\substack{x=x_0 \\ y=m_i}} = \left[\sum_{s=0}^{M} \frac{1}{s!} \frac{\partial^{s+1} P_y(x)}{\partial y^{s+1}} z_j^s\right]_{\substack{x=x_0 \\ y=N}} \tag{11}$$

$$\left.\frac{\partial^2 P_y(x)}{\partial x \partial y}\right|_{\substack{x=x_0 \\ y=m_i}} = \left[\sum_{s=0}^{M} \frac{1}{s!} \frac{\partial^{s+2} P_y(x)}{\partial x \partial y^{s+1}} z_j^s\right]_{\substack{x=x_0 \\ y=N}}$$

where $N = j$ in one case and $j + \frac{1}{2}$ in the other. The expressions appearing in the right hand members of (11) are all obtainable by means of the methods of reference 1. When the computed value of z_j is close to zero, $M = 1$ suffices in the sums appearing in (11). However, for values of z_j close to $\frac{1}{4}$ (which, in the final analysis, is the largest value one must consider for z_j) a greater number of terms in these sums should be used, i.e. $M \geqq 2$.

Upon applying this method to the case of $\theta_0 = 165°$ the results indicate that the two values of the zeros obtained bound the true value. For example, the values of m_4 computed by this technique were found to be 3.8900 and 3.8597. If bounds on m_4 are obtained using the hypergeometric series method discussed in reference 2, one finds that $3.86 < m_4 < 3.89$. A similar situation occurred in connection with the n_i's of reference 1.

In the following table the results obtained for the $\theta_0 = 165°$ case are given

$x_0 = \cos 165°$

i	m_i		$\int_{x_0}^{1} [P_{m_i}^1(x)]^2\, dx$		
			$M = 1$		$M = 2$
	$j + z_j$ Method	$j + \frac{1}{2} + z_j$ Method	$j + z_j$ Method	$j + \frac{1}{2} + z_j$ Method	$j + \frac{1}{2} + z_j$ Method
1	.9673*	—	1.3640*	—	—
2	1.9198*	1.8850	2.4827*	3.3509	—
3	2.8894*	2.8572	3.5179*	4.7478	—
4	3.8900*	3.8597	4.4494*	6.3861	—
5	4.9180*	4.8872	5.2909*	8.6650	—
6	5.9657*	5.9336	6.1148*	12.076	—
7	7.0264*	7.0900	7.0519*	12.754	—
8	8.0940*	8.1436	8.2651*	12.847	—
9	9.1638*	9.2013	9.9351*	12.670	—
10	10.233*	10.264	12.270	12.394	10.089*
11	11.302	11.333*	15.608	12.206	11.539*
12	12.373	12.410*	20.505	12.294	12.032*
13	13.450	13.492*	27.915	12.887	12.830*
14	14.626	14.576*	23.618	14.204	13.710*
15	15.694	15.658*	21.489	16.422	14.108*

* Indicates preferred value.

and some of the integrals computed for both $M = 1$ and $M = 2$ are presented for comparison purposes. In computing the integrals with $M = 1$, this process yields values which are larger than the true values; in fact the envelope of the resulting curves would more accurately define the true values. It has been found that, due to the periodic changes in the total error, in many cases it is only necessary to compute a few terms with increased accuracy ($M = 2$), while the $M = 1$ cases suffice for most engineering computations.

There has been no attempt to smooth out the entries which appear in the table since such an attempt would imply information relative to the values not presented by the method itself. The authors would again like to point out that this method is not as accurate as certain other known processes. However, it does yield results much more economically than any other known standard technique.

It also should be pointed out that large scale digital machines might be applied to obtaining tables with a large number of entries. In the usual electromagnetic theory scattering problem, one usually uses a summing technique which results in the determination of which terms are dominant and then one only resorts to computing a few of these Legendre Functions and their normalizing integrals. When complete tables come into existence, they undoubtedly will have been obtained by large scale digital machines. In the intervening period, an approximation technique such as this is necessary. It is clear that the method presented in this note can be applied to other functions as well as the Legendre Functions.

REFERENCES

1. Siegel, K. M., Crispin, J. W., Kleinman, R. E. and Hunter, H. E., The Zeros of $P^1_{n+\frac{1}{2}}(x_0)$ of Non-Integral Degree. Journal of Mathematics and Physics (October, 1952).
2. Siegel, K. M., Brown, D. M., Hunter, H. E., Alperin, H. A. and Quillen, C. W., UMM-82: Studies in Radar Cross-Sections—II, The Zeros of the Associated Legendre Functions of Non-Integral Degree. Willow Run Research Center, University of Michigan, April, 1951.
3. Magnus, W. and Oberhettinger, F., Formulas and Theorems for the Special Functions of Mathematical Physics. Chelsea Publishing Company, 1949.
4. Siegel, K. M. and Alperin, H. A., UMM-87: Studies in Radar Cross-Sections—III, Scattering by a Cone. Willow Run Research Center, University of Michigan, January, 1952.

Willow Run Research Center
University of Michigan

(Received December 12, 1952)

RADIX TABLE FOR OBTAINING HYPERBOLIC AND INVERSE HYPERBOLIC FUNCTIONS TO MANY PLACES

By Herbert E. Salzer

Introduction. Hyperbolic and inverse hyperbolic functions have almost the same importance as exponential and logarithmic functions for the needs of mathematicians, statisticians and physical scientists. Key values of these functions to many places are handy in numerous situations. For instance, there are recurrence formulas where very many decimal places are lost in a few applications. Then in order to attain three or four place accuracy (which might be all that is needed for physical purposes) one must begin the recursion scheme with what might appear to be a fantastically large number of significant figures. Thus one example where an eighteen place table of $\tanh^{-1} x$ would be useful to physicists is in the generation by recursion of the functions $Q_n(x)$, the Legendre functions of the second kind which are vital in potential problems. The recurrence formula

$$nQ_n(x) + (n - 1)Q_{n-2}(x) - (2n - 1)xQ_{n-1}(x) = 0$$

might be used to generate $Q_n(x)$ for large x and n from $Q_0(x) = \tanh^{-1} x$ and $Q_1(x) = x \tanh^{-1} x - 1$. But the error is multiplied by around $2x$ at each step. Thus if x were around 50 and we wanted to obtain $Q_8(x)$ by seven successive applications of the recurrence formula, losing possibly two places at each application, by the time only $Q_8(x)$ is reached we might have only two decimals left out of the original eighteen.

The purpose of this article is to give in a few pages the means of dispensing with the equivalent of more extensive tables of hyperbolic and inverse hyperbolic tangents (of course, at the expense of some additional work in using it). The method and table here can be adapted to obtain fewer than eighteen places, using a smaller number of steps and decimal places.

In a sense this article is a companion one to a similar radix table for generating tangents and arctangents.[1] Unlike radix schemes for calculating logarithms and antilogarithms which have been in use for over a century, radix schemes for circular and hyperbolic functions necessarily involve division, so that they are not feasible beyond twenty-place tables.

The scheme for the hyperbolic functions is not strictly parallel to that for the circular functions, the tables in the former having to be supplemented by a few additional values of $\tanh x$ and $(\frac{1}{2})\log_e x$. The reason for this difference is due to the presence of a minus sign in the denominator of the subtraction formula for $\tanh x$, which affects the convergence of the algorithm for certain arguments (unlike the case of $\tan x$ with a plus sign in the denominator).

This present table enables one to calculate hyperbolic tangents and the inverse of hyperbolic tangents, and thence all the other hyperbolic and inverse hyperbolic functions, in almost every case to at least 18*D*. The method employs two

[1] H. E. Salzer, "Radix Table for Trigonometric Functions and Their Inverses to High Accuracy," *Math. Tables and Other Aids to Comput.*, vol. V, Jan. 1951, pp. 9–11.

algorithms, both of which follow from repeated application of the addition formula for tanh x, and a radix table of Arc tanh $(m\ 10^{-n})$, $m = 1(1)9$, $n = 1(1)6$, together with supplementary tables of tanh A, $A = 1(1)24$, $\frac{1}{2}\log I$, $I = 2(1)10$, and $\frac{1}{2}\log J$, $J = 1.2(.2)1.8$, all given to $20D$. The division operations may be done upon an ordinary 10-bank desk calculator as 19 by 19 divisions, which has the advantage of not requiring any intermediate written work for any division. However, if one is willing to perform 20 by 20 divisions, in almost every instance one will obtain tanh x or Arc tanh t to $19D$.

1. To obtain tanh x to 18 decimals: If $x < 1$, subtract from x the closest tabular value of Arc tanh a_i, which leaves a remainder $\pm x_i$; from x_i subtract the closest value of Arc tanh a_j, $j \geqq i$, $i = 1$, $j > i$, $i \neq 1$, which leaves a remainder $\pm x_j$; from x_j subtract the closest value of Arc tanh a_k, $k > j$, which leaves a remainder $\pm x_k$, and so on. Continue in this way until the remainder $\pm x_0$, where $x_0 \leqq \frac{1}{2}\cdot 10^{-6}$ is reached, after no more than seven subtractions. Thus (1) $x = \text{Arc tanh } a_i \pm (\text{Arc tanh } a_j \pm (\text{Arc tanh } a_k \pm \cdots \pm (\text{Arc tanh } a_0 \pm x_0)) \cdots))$, where the right member will not have more than eight terms after the removal of the parentheses. Assume that x_0 is the eighth term, and if there are fewer than seven Arc tanh terms, bring the number up to seven by including $+\text{Arc tanh } 0$ the requisite number of times until Arc tanh a_0 is reached. In (1), after the removal of parentheses, set $a_r' = a_r$ when Arc tanh a_r is preceded by a $+$ sign, and set $a_r' = -a_r$ when Arc tanh a_r is preceded by a $-$ sign, where $r = j, k, \cdots, 0$; and also set $x_0' = \pm x_0$. Evaluate

$$a = a_i + a_j',\ b = a_k' + a_l',\ c = a_m' + a_n',\ d = a_0' + x_0';$$

$$e = 1 + a_i a_j',\ f = 1 + a_k' a_l',\ g = 1 + a_m' a_n',\ h = 1 + a_0' x_0';$$

$$A_1 = af + be,\ A_2 = ab + ef,\ B_1 = ch + dg,\ B_2 = cd + gh.$$

Then $t = \tanh x \doteq (A_1B_2 + A_2B_1)/(A_1B_1 + A_2B_2)$.

If $24 > x \geqq 1$, and A denotes the largest integer $\leqq x$, let $x = A + x'$, $0 \leqq x' < 1$. First find tanh x' and then calculate

$$t = \tanh x = (\tanh A + \tanh x')/(1 + \tanh A \tanh x')$$

Other functions are given by

$$\sinh x = t/\sqrt{1 - t^2},\ \cosh x = 1/\sqrt{1 - t^2},\ \coth x = 1/t,$$

$$\operatorname{sech} x = \sqrt{1 - t^2},\ \operatorname{cosech} x = \sqrt{1 - t^2}/t.$$

2. To obtain Arc tanh t to 18 decimals:[2] If $T = (1 + t)/(1 - t) > 1.1$, calculate $T' = T/10^m\ I^{\pm 1}\ J^{\pm 1}$, where I is an integer ranging from 2 to 9, J is one of the four quantities 1.2, 1.4, 1.6 and 1.8, the divisors 10^m, $I^{\pm 1}$, and $J^{\pm 1}$ being

[2] An exception to this estimate occurs whenever $1 - t$ has fewer than 19 significant figures. Thus if $1 - t$ has S significant figures, Arc tanh t can be calculated to $S - 1$ decimals.

Table of Arc tanh a_i

a_1	Arc tanh a_1	a_4	Arc tanh a_4
.1	.10033 53477 31075 58064	.0001	.00010 00000 00333 33334
.2	.20273 25540 54082 19099	.0002	.00020 00000 02666 66673
.3	.30951 96042 03111 71547	.0003	.00030 00000 09000 00049
.4	.42364 89301 93601 80686	.0004	.00040 00000 21333 33538
.5	.54930 61443 34054 84570	.0005	.00050 00000 41666 67292
.6	.69314 71805 59945 30942	.0006	.00060 00000 72000 01555
.7	.86730 05276 94053 19443	.0007	.00070 00001 14333 36695
.8	1.09861 22886 68109 69140	.0008	.00080 00001 70666 73220
.9	1.47221 94895 83220 23000	.0009	.00090 00002 43000 11810
a_2	Arc tanh a_2	a_5	Arc tanh a_5
.01	.01000 03333 53334 76202	.00001	.00001 00000 00000 33333
.02	.02000 26673 06849 58072	.00002	.00002 00000 00002 66667
.03	.03000 90048 63126 47433	.00003	.00003 00000 00009 00000
.04	.04002 13538 36768 21291	.00004	.00004 00000 00021 33333
.05	.05004 17292 78491 26825	.00005	.00005 00000 00041 66667
.06	.06007 21559 21031 62366	.00006	.00006 00000 00072 00000
.07	.07011 46706 54325 11799	.00007	.00007 00000 00114 33333
.08	.08017 13250 37589 69169	.00008	.00008 00000 00170 66667
.09	.09024 41878 56146 82961	.00009	.00009 00000 00243 00000
a_3	Arc tanh a_3	a_6	Arc tanh a_6
.001	.00100 00003 33333 53333	.000001	.00000 10000 00000 00033
.002	.00200 00026 66673 06668	.000002	.00000 20000 00000 00267
.003	.00300 00090 00048 60031	.000003	.00000 30000 00000 00900
.004	.00400 00213 33538 13567	.000004	.00000 40000 00000 02133
.005	.00500 00416 67291 67783	.000005	.00000 50000 00000 04167
.006	.00600 00720 01555 23999	.000006	.00000 60000 00000 07200
.007	.00700 01143 36694 85099	.000007	.00000 70000 00000 11433
.008	.00800 01706 73220 56627	.000008	.00000 80000 00000 17067
.009	.00900 02430 11810 48332	.000009	.00000 90000 00000 24300

chosen so as to bring T' as close to 1 as possible. From $T' = (1 + t')/(1 - t')$, $t' = (T' - 1)/(T' + 1) = \pm|\, t'\,|$. Then

(2) Arc tanh $t = m(\frac{1}{2} \log 10) \pm \frac{1}{2} \log I \pm \frac{1}{2} \log J \pm$ Arc tanh $|\, t'\,|$.

To find Arc tanh $|\, t'\,|$, evaluate in turn

$$t_1 = (|\, t'\,| - a_i)/(1 - a_i\,|\, t'\,|) = \pm\,|\, t_1\,|\,,$$

where a_i is the closest single-digit decimal to $|\, t'\,|$,

$$t_2 = (|\, t_1\,| - a_j)/(1 - a_j\,|\, t_1\,|) = \pm|\, t_2\,|,$$

where a_j is the closest single-digit decimal to $|\, t_1\,|$, and so on, using

$$t_p = (|\, t_{p-1}\,| - a_q)/(1 - a_q\,|\, t_{p-1}\,|) = \pm|\, t_p\,|,$$

where a_q is the closest single-digit decimal to $|\, t_{p-1}\,|$, until t_r, where $|\, t_r\,| \leqq \frac{1}{2} \cdot 10^{-6}$, is reached after no more than five such steps. Then

Tables of tanh A, $\frac{1}{2}\log I$, *and* $\frac{1}{2}\log J$

A	Tanh A	I	$\frac{1}{2}\log I$
1	.76159 41559 55764 88812	10	1.15129 25464 97022 84201
2	.96402 75800 75816 88395		
3	.99505 47536 86730 45133		
4	.99932 92997 39067 04379	2	.34657 35902 79972 65471
5	.99990 92042 62595 13121	3	.54930 61443 34054 84570
		4	.69314 71805 59945 30942
6	.99998 77116 50795 57056	5	.80471 89562 17050 18730
7	.99999 83369 43944 67174	6	.89587 97346 14027 50041
8	.99999 97749 29675 88981	7	.97295 50745 27656 65255
9	.99999 99695 40040 97448	8	1.03972 07708 39917 96413
10	.99999 99958 77692 76362	9	1.09861 22886 68109 69140
11	.99999 99994 42106 38158	J	$\frac{1}{2}\operatorname{Log} J$
12	.99999 99999 24497 30912		
13	.99999 99999 89781 82194	1.2	.09116 07783 96977 31311
14	.99999 99999 98617 11998	1.4	.16823 61183 10606 46525
15	.99999 99999 99812 84754	1.6	.23500 18146 22867 77683
		1.8	.29389 33324 51059 50409
16	.99999 99999 99974 67167		
17	.99999 99999 99996 57218		
18	.99999 99999 99999 53610		
19	.99999 99999 99999 93722		
20	.99999 99999 99999 99150		
21	.99999 99999 99999 99885		
22	.99999 99999 99999 99984		
23	.99999 99999 99999 99998		
24	1.00000 00000 00000 00000		

(3) Arc tanh $|t'| \doteq$ Arc tanh $a_i \pm$ (Arc tanh $a_j \pm$ (Arc tanh $a_k \pm \cdots \pm$ (Arc tanh $a_s \pm |t_r|)) \cdots$)),

where the sign of $\pm$(Arc tanh a_q is the same as the sign of $t_{p-1} = \pm|t_{p-1}|$.

Other functions are given by

Arc coth t = Arc tanh $(1/t)$, Arc sinh t = Arc tanh $(t/\sqrt{1+t^2})$,

Arc cosh t = Arc tanh $(\sqrt{t^2-1}/t)$, Arc sech t = Arc tanh $\sqrt{1-t^2}$,

Arc cosech t = Arc tanh $(1/\sqrt{1+t^2})$.

Examples

A. *To obtain* tanh x, *for* x = 2.74531 68104 55937 84126:

Here $x = A + x'$, $A = 2$, and $x' = .74531\ 68104\ 55937\ 84126$.

For x', the closest tabular value of Arc tanh a_i is for $a_1 = .6$,

$x_1 = x' -$ Arc tanh .6 = .74531 68104 55937 84126

$-$ .69314 71805 59945 30942 = .05216 96298 95992 53184.

For x_1, the closest tabular value of Arc tanh a_j is for $a_2 = .05$,

$x_2 = x_1 - \text{Arc tanh } .05 = .05216\ 96298\ 95992\ 53184$

$- .05004\ 17292\ 78491\ 26825 = .00212\ 79006\ 17501\ 26359.$

For x_2, the closest tabular value of Arc tanh a_k is for $a_3 = .002$,

$x_3 = x_2 - \text{Arc tanh } .002 = .00212\ 79006\ 17501\ 26359$

$- .00200\ 00026\ 66673\ 06668 = .00012\ 78979\ 50828\ 19691.$

For x_3, the closest tabular value of Arc tanh a_l is for $a_4 = .0001$,

$x_4 = x_3 - \text{Arc tanh } .0001 = .00012\ 78979\ 50828\ 19691$

$- .00010\ 00000\ 00333\ 33334 = .00002\ 78979\ 50494\ 86357.$

For x_4, the closest tabular value of Arc tanh a_m is for $a_5 = .00003$,

$-x_5 = x_4 - \text{Arc tanh } .00003 = .00002\ 78979\ 50494\ 86357$

$- .00003\ 00000\ 00009\ 00000 = -.00000\ 21020\ 49514\ 13643.$

For x_5, the closest tabular value of Arc tanh a_n is for $a_6 = .000002$,

$x_6 = x_5 - \text{Arc tanh } .000002 = .00000\ 21020\ 49514\ 13643$

$- .00000\ 20000\ 00000\ 00267 = .00000\ 01020\ 49514\ 13376.$

From (1), after removal of parentheses, $a_i = .6$, $a_j' = .05$, $a_k' = .002$, $a_l' = .0001$, $a_m' = .00003$, $a_n' = -.000002$, $a_0' = 0$, and $x_0' = -.00000\ 01020\ 49514\ 13376$.

$a = a_i + a_j' = .65$, $b = a_k' + a_l' = .0021$, $c = a_m' + a_n' = .000028$, $d = a_0' + x_0' = -.00000\ 01020\ 49514\ 13376$, $e = 1 + a_i a_j' = 1.03$, $f = 1 + a_k' a_l' = 1.00000\ 02$, $g = 1 + a_m' a_n' = .99999\ 99999\ 4$, $h = 1 + a_0' x_0' = 1$;

$A_1 = af + be = .65216\ 3130$, $A_2 = ab + ef = 1.03136\ 5206$,
$B_1 = ch + dg = .00002\ 78979\ 50485\ 87236$,
$B_2 = cd + gh = .99999\ 99999\ 37142\ 61360$;
$A_1B_2 = .65216\ 31299\ 59006\ 73014$, $A_2B_1 = .00002\ 87729\ 75449\ 83955$,
$A_1B_1 = .00001\ 81940\ 14709\ 45154$, $A_2B_2 = 1.03136\ 52059\ 35171\ 0787$,
$A_1B_2 + A_2B_1 = .65219\ 19029\ 34456\ 56969$,
$A_1B_1 + A_2B_2 = 1.03138\ 33999\ 49880\ 5302$.

$$\tanh x' \doteq \frac{A_1B_2 + A_2B_1}{A_1B_1 + A_2B_2} = .63234\ 67131\ 29326\ 5445,$$

$$t = \tanh x = \frac{\tanh 2 + \tanh x'}{1 + \tanh 2 \tanh x'} \doteq (.96402\ 75800\ 75816\ 88395$$

$$+ .63234\ 67131\ 29326\ 5445)/(1 + .60959\ 96716\ 26961\ 4531)$$

$$= \frac{1.59637\ 42932\ 05143\ 4285}{1.60959\ 96716\ 26961\ 4531} = .99178\ 34361\ 82954\ 7392,$$

which is correct to about 1 unit in the 19th decimal place.

B. *To obtain* Arc tanh t, *for* t = .93724 53698 44713 27658:

$$T = \frac{1+t}{1-t} = \frac{1.93724\ 53698\ 44713\ 2766}{.06275\ 46301\ 55286\ 72342},$$

or T = 30.87015 83461 01116 93(2).

Here $m = 1$, $I = 3$, and a divisor J is not required.

$$T/30 = T' = 1.02900\ 52782\ 03370\ 5644.$$

$$t' = \frac{T'-1}{T'+1} = \frac{.02900\ 52782\ 03370\ 5644}{2.02900\ 52782\ 03370\ 564},$$

$$\text{or } t' = .01429\ 53192\ 45819\ 7821 = |t'|.$$

Since

$$a_i = a_2 = .01,\ t_1 = (|t'| - .01)/(1 - .01\,|t'|)$$

$$= \frac{.00429\ 53192\ 45819\ 7821}{.99985\ 70468\ 07541\ 8022} = .00429\ 59333\ 63208\ 6404(6) = |t_1|.$$

From $a_j = a_3 = .004$, $t_2 = (|t_1| - .004)/(1 - .004\,|t_1|)$

$$= \frac{.00029\ 59333\ 63208\ 6404(6)}{.99998\ 28162\ 66547\ \ 165} = .00029\ 59384\ 48536\ 0585(5) = |t_2|.$$

From $a_k = a_4 = .0003$, $t_3 = (|t_2| - .0003)/(1 - .0003\,|t_2|)$

$$= -\frac{.00000\ 40615\ 51463\ 9414(5)}{.99999\ 99112\ 18465\ \ 44} = -.00000\ 40615\ 51824\ 5322(5)$$

$$= -|t_3|.$$

From $a_l = a_5 = .000004$, $t_4 = (|t_3| - .000004)/(1 - .000004\,|t_3|)$

$$= \frac{.00000\ 00615\ 51824\ 5322(5)}{.99999\ 99999\ 83754} = .00000\ 00615\ 51824\ 5332(5) = |t_4|.$$

From (3), Arc tanh $|t'| \doteq$ Arc tanh .01 + Arc tanh .004 + Arc tanh .0003 − Arc tanh .000004 − .00000 00615 51824 5332(5), or

$$\text{Arc tanh } |t'| \doteq .01429\ 62931\ 44048\ 3436(0).$$

$$\text{Arc tanh } t = \tfrac{1}{2}\log 10 + \tfrac{1}{2}\log 3 + \text{Arc tanh } |t'|$$

$$\doteq 1.15129\ 25464\ 97022\ 84201 + .54930\ 61443\ 34054\ 84570$$

$$+ .01429\ 62931\ 44048\ 3436(0) = 1.71489\ 49839\ 75126\ 0313(1),$$

which is correct to about 1 unit in the 19th decimal place.

NATIONAL BUREAU OF STANDARDS

(Received September 22, 1952)

A NOTE ON THE NETWORK POSTULATES*

By Maynard G. Arsove

1. The classical network postulates are given in the form of Kirchoff's laws:

I. The sum of the currents at a junction point is zero;

II. The sum of the potential differences about any closed path of branches is zero.

These are, of course, just special cases of the general physical laws of conservation of matter and conservation of energy. However, as the laws are stated, there is a striking lack of parallelism between their formulations. The first law is expressed in terms of the individual junction points, or nodes, while the second law is expressed not in terms of the individual branches, but in terms of closed paths of branches. From the point of view that the linear graph of a network is completely determined by the incidence relations between its nodes and branches, the manner in which law I is presented may be regarded as more fundamental than that of law II.

What we propose to do is to maintain law I but replace law II by a version of the law of conservation of energy making no explicit reference to paths of branches. As will be seen, the resulting law can be interpreted geometrically in terms of orthogonal subspaces of a certain vector space. Moreover, the network definition which we give is readily shown to be invariant under changes of basis. This invariance property, which is essential for the definition to be meaningful, seems to be accepted without proof in the usual discussions of Kirchoff's laws. We establish, finally, the equivalence of our definition with the classical one.

2. The formulation of the network problem. We start with a linear graph G having n branches and m nodes and let $\bar{G}$ denote the set of all $2n$ oriented branches that can be formed from branches of G. It will be assumed further that the two end points of any branch of G are distinct.

A complex-valued function f with domain $\bar{G}$ will be called *antisymmetrical* provided $f(X_r) = -f(X)$, where X is any oriented branch and X_r the oppositely oriented branch. Similarly, a complex-valued function g defined on the Cartesian product $\bar{G} \times \bar{G}$ will be called *antisymmetrical* provided

$$g(X_r, Y) = g(X, Y_r) = -g(X, Y),$$

where X and Y are any oriented branches and X_r and Y_r are the corresponding oppositely oriented branches.

Definition. By a *network* we mean an ordered triple $N = (G, E, Z)$, where G is a linear graph no branch of which has coincident end points, E is an antisymmetrical function on $\bar{G}$, and Z is an antisymmetrical function on $\bar{G} \times \bar{G}$. E is called the *emf* and Z the *impedance* of N.

The reason for excluding from consideration linear graphs having branches

* Based on a portion of the author's thesis, *Algebraic Theory of Linear Networks*, submitted in partial fulfillment of the requirements for the degree of Master of Science at Brown University (1948).

which close on themselves is a technical one: whenever a branch has distinct end points, an orientation of that branch is uniquely determined by an ordered pair of its end points. Moreover, from the physical viewpoint this restriction entails no real loss of generality, since any given node can be regarded as a pair of nodes joined by a direct connection.

The terminology *basis in* G will be used to denote an ordered pair (P, B) with P a sequential ordering of the m nodes as $P_1, P_2, \cdots, P_m$ and B a sequential ordering of n oriented branches as $B_1, B_2, \cdots, B_n$ and having the property that none of the oriented branches is a reverse of any other. Once a basis in G has been prescribed, the antisymmetrical function E can be expressed uniquely as an n-dimensional column vector whose k^{th} component is the complex number $E(B_k)$. Similarly, Z appears uniquely as the $n \times n$ matrix whose elements are the complex numbers $Z(B_j, B_k)$. We shall effect an economy of notation at no serious expense in clarity, by using E and Z to denote both the functions and the corresponding vectors and matrices in a specified basis.

Any basis in G determines an *orientation matrix*, or *node matrix*, F defined as

$$F_{jk} = \begin{cases} -1 \text{ if } P_j \text{ is the initial point of } B_k \\ 0 \text{ if } P_j \text{ is not an end point of } B_k \\ +1 \text{ if } P_j \text{ is the terminal point of } B_k . \end{cases}^{1}$$

Since it does not seem possible to define the network problem in a basis-free manner, we shall formulate the problem in terms of a particular basis and then prove that the resulting definition is unaffected by a change of basis. In what follows an asterisk will be used to denote the transpose of a matrix.

DEFINITION. Given a network $N = (G, E, Z)$, the *linear network problem* for N consists of determining two antisymmetrical functions V (the *voltage*) and I (the *current*) on $\bar{G}$ such that the following properties hold:

1) relative to some basis β of G, the vectors V and I satisfy the matrix equation

$$V = E - ZI; \tag{2.1}$$

2) for F the node matrix determined by β

$$FI = 0; \tag{2.2}$$

3) all vectors I satisfying (2.2) also satisfy

$$V^*I = 0. \tag{2.3}$$

It is apparent that condition (2.2) is precisely Kirchoff's first law, since the j^{th} component of FI is the sum of those $I(b)$ for which b is an oriented branch with terminal point P_j. To show that condition (2.3) expresses the law of conservation of energy, we combine it with (2.1) to obtain

$$E^*I = I^*ZI \tag{2.4}$$

[1] Non-coincidence of the end points of each branch of G is essential here.

for all admissible current vectors I. Equation (2.4), however, simply asserts that the *instantaneous generated power* E^*I must equal the *instantaneous absorbed power* I^*ZI.

Let us digress for a moment to discuss the geometrical import of postulates (2.2) and (2.3). The vectors with which we have been dealing are all n-tuples of complex numbers; that is to say, they lie in n-dimensional complex Euclidean space K^n. Those vectors x of K^n satisfying $Fx = 0$ comprise a linear subspace S known as the *null space* of F. Hence, condition (2.2) is the same as

$$I \in S. \tag{2.2'}$$

The inner product of two vectors x and y in K^n is given by $x^*\tilde{y}$, where $\tilde{y}$ denotes the complex conjugate of y. Now, since the elements of F are real numbers, a vector I is in S if and only if its real and imaginary parts are vectors in S. It follows that $V^*I = 0$ for all $I \in S$ is equivalent to $V^*\tilde{I} = 0$ for all $I \in S$. Geometrically, this states that V is orthogonal to all vectors in S. We can therefore express condition (2.3) concisely as

$$V \in S^{\perp}, \tag{2.3'}$$

where $S^{\perp}$, the *orthogonal complement* of S, is the subspace consisting of those vectors orthogonal to all vectors in S.

In order to establish invariance of our formulation of the network problem under a change of basis from β to β', let us first find the transformations on the various entities involved in these formulations. The new ordering P' of the nodes is some permutation XP of the original ordering, and the new ordering $B' = YB$ of the oriented branches is the result of permuting the original oriented branches and perhaps reversing some of them. As far as the corresponding operations on antisymmetrical functions (considered as vectors and matrices) are concerned, we can regard X as a permutation matrix and Y as the product of a permutation matrix and a diagonal matrix of elements ± 1. It is then readily seen that $E' = YE$, $I' = YI$, $V' = YV$, $Z' = YZY^*$, and $F' = XFY^*$. Observing that the matrices X and Y are orthogonal, we infer that $V' = E' - Z'I'$ is equivalent to $V = E - ZI$, that $F'I' = 0$ is equivalent to $FI = 0$, and that $V'^*I' = 0$ is equivalent to $V^*I = 0$. Hence, the formulation of the network problem is independent of the choice of basis.

A network is said to be *symmetrical* provided that in some basis its impedance appears as a symmetrical matrix. That this definition is an appropriate one follows from the equation of transformation of the impedance matrix, since it is obvious that symmetry of the impedance matrix in one basis implies its symmetry in all bases.

3. Solution of the network problem. Loops. If the rank ρ of the node matrix F is n, then its null space S reduces to the origin, and the given network N has the unique trivial solution $I = 0$, $V = E$. When $\rho < n$, the null space S has dimension $n - \rho$, so that we can solve $FI = 0$ as $I = Ai$, where A is an $n \times (n - \rho)$ matrix and the vector i is formed of any $n - \rho$ linearly independent

components of I. Since then all complex-valued vectors i satisfy $FAi = 0$, the energy postulate (2.3) becomes

$$V^*Ai = 0 \quad \text{for } \textit{all } i,$$

which is equivalent to

$$A^*V = 0.$$

Application of this result to (2.1) gives rise to

$$A^*E = A^*ZAi. \tag{3.1}$$

Hence, the network problem for N has a unique solution if and only if A^*ZA is non-singular, the solution then being given by

$$\begin{cases} I = A(A^*ZA)^{-1}A^*E, \\ V = E - ZA(A^*ZA)^{-1}A^*E. \end{cases} \tag{3.2}$$

It is clear from this that the solution, as well as the statement, of the network problem can be given without reference to closed paths of branches.

By a *loop* of the graph G is meant a connected subgraph L of G such that each node of L is incident with exactly two branches of L. An *oriented loop* of G is an oriented linear graph for which the associated graph is a loop and for which each of its nodes is the initial point of one of the branches with which it is incident and the terminal point of the other branch with which it is incident. In any given basis β an oriented loop Λ can be represented by an n-dimensional column vector λ whose k^{th} component is

$$\lambda_k = \begin{cases} -1 \text{ if } B_k \text{ reversed is in } \Lambda \\ 0 \text{ if neither } B_k \text{ nor } B_k \text{ reversed is in } \Lambda \\ +1 \text{ if } B_k \text{ is in } \Lambda. \end{cases}$$

An important consequence of topological considerations is that the columns of A form a set of $n - \rho$ linearly independent loop vectors, and conversely that the general solution of $FI = 0$ can be expressed as $I = \tilde{A}\tilde{\imath}$, where $\tilde{A}$ is an $n \times (n - \rho)$ matrix whose columns are linearly independent loop vectors and $\tilde{\imath}$ is an arbitrary $(n - \rho)$-dimensional complex-valued vector. An independent proof of these assertions presents no special difficulties, but we omit it here.[2]

On the basis of this theorem it is easy to verify that (3.1) is, in matrix form, just the familiar "loop equations" for N corresponding to the $n-\rho$ loops comprising the columns of A. Furthermore, we have the result that for λ any loop vector and V the voltage vector in the same basis $\lambda^*V = 0$: the sum of the voltages around any loop is zero. The proof consists simply of the fact that a matrix $\tilde{A}$ can be constructed with λ as a column. Kirchoff's second law is thus a consequence of our postulates. Conversely, Kirchoff's second law implies $A^*V = 0$ and thereby the energy condition (2.3).

UNIVERSITY OF WASHINGTON

(Received December 1, 1952)

[2] See, for example, the author's thesis.

A NOTE ON CONTACT TRANSFORMATIONS

By H. D. Block

The purpose of this note is to dispel a rather common misconception by means of an elementary example.

Let E denote Euclidean $2n$ space whose points $(q_1, \cdots, q_n, p_1, \cdots, p_n)$ are also denoted by (q, p).

If: (a) $P_i(q, p), Q_i(q, p)(i = 1, 2, \cdots, n)$ are C^1 in a non-null open set $E' \subseteqq E$, (b) the $2n$ equations

$$\bar{q}_i = Q_i(q, p), \quad \bar{p}_i = P_i(q, p) \qquad (i = 1, 2, \cdots, n) \quad (1)$$

define a one-to-one mapping of E' onto an open set $E'' \subseteqq E$; and if

(c) there exists an $F(q, p) \in C^1$ in E' such that

$$\sum_{i=1}^{n} P_i \, dQ_i - \sum_{i=1}^{n} p_i \, dq_i = dF,$$

then the mapping defined by (1) is called a *contact transformation*. If now $S(q, \bar{p}) \in C^2$ and $| \partial^2 S/\partial q_i \partial \bar{p}_j | \neq 0$ in $E' \times E''$ then it can be shown that the equations

$$p_i = (\partial S/\partial q_i), \quad \bar{q}_i = (\partial S/\partial \bar{p}_i) \qquad (i = 1, 2, \cdots, n) \quad (2)$$

can be reduced to the form (1) and that the mapping thus determined is a contact transformation. Furthermore, the variables $\{q, \bar{p}\}$ are here said to be an *independent set* since one can determine uniquely the remaining coordinates $\{\bar{q}, p\}$ in terms of them. Conversely any contact transformation for which $\{q, \bar{p}\}$ are an independent set can be obtained in this way from a suitable function $S(q, \bar{p})$.*

Similar results are obtained if one starts instead with suitable functions $S(q, \bar{q})$, $S(p, \bar{q})$, or $S(p, \bar{p})$. The necessary changes having been made in (2), one finds all those (and only those) contact transformations in which $\{q, \bar{q}\}$, $\{p, \bar{q}\}$, $\{p, \bar{p}\}$ are, respectively, independant sets.

Now it is believed by some mathematicians and physicists that every contact transformation can be generated in at least one of these four ways.† In order to show that this is not the case it is sufficient to exhibit a contact transformation for which none of the four sets mentioned is an independent set. Now the transformation defined by $\bar{q}_1 = q_1$, $\bar{q}_2 = p_2$, $\bar{p}_1 = p_1$, $\bar{p}_2 = -q_2$ is readily verified to be a contact transformation with $F = -p_2 q_2$. On the other hand $\{q, \bar{p}\}$ does not determine $\bar{q}_2$ or p_2, $\{q, \bar{q}\}$ does not determine $\bar{p}_1$ or p_1, $\{p, \bar{q}\}$ does not determine q_2 or $\bar{p}_2$, and $\{p, \bar{p}\}$ does not determine q_1 or $\bar{q}_1$, so that this transformation could not possibly be generated in the manner suggested.

* See, e.g., Osgood "Mechanics," MacMillan Co. 1937, p. 391.

† See, e.g., MacMillan "Theoretical Mechanics: Dynamics of Rigid Bodies," McGraw Hill Book Co. 1936, p. 364; or Goldstein "Classical Mechanics," Addison Wesley, 1950, p. 240 and p. 248 (footnote).

A somewhat more general contact transformation is defined by the equations:

$$\bar{q}_1 = aq_1\,,\ \bar{q}_2 = bp_2\,,\ \bar{p}_1 = cq_1 + (1/a)p_1 + dp_2\,,\ \bar{p}_2 = (ad/b)q_1 - (1/b)q_2 + ep_2\,,$$

where a, b, c, d, e, are real constants and $ab \neq 0$. If $e \neq 0$ this is generated by an $S(q, \bar{p})$, if $c \neq 0$ it is generated by an $S(p, \bar{p})$, while if $c = 0$ and $e = 0$ it cannot be generated by any of the four methods.

As far as the writer is aware, the existence of contact transformations which cannot be generated by any of the four generating functions has not been previously noted in the literature.

IOWA STATE COLLEGE

(Received September 15, 1952)

"MONTE CARLO" METHODS FOR THE ITERATION OF LINEAR OPERATORS

By J. H. Curtiss

1. Introduction. A purely formal description of the type of problem to be dealt with in this paper is as follows. Let $c = c(x)$ and $u_0 = u_0(x)$ be real-valued functions defined on a coordinate space R, which may be multidimensional. Let $L = L(f)$ be a linear transformation defined on the space of all real-valued functions f whose arguments belong to R. Required, to calculate the sequence of functions $u_1, u_2, \cdots$, defined by the recursion formula

$$u_{N+1} = L(u_N) + c, \qquad N = 0, 1, 2, \cdots \tag{1.1}$$

This problem arises in many contexts in both pure and applied mathematics. Much attention has recently been focused on the numerical aspects of it by the nuclear physicists, because the recursion formula (1.1) is obtained when time-dependent diffusion and transport problems are formulated in a discrete form. The methods to be discussed here for estimating the solution are an outgrowth of a novel stochastic attack suggested during the late war by von Neumann and Ulam in connection with diffusion problems.[1] Their idea was to bypass the mathematical formulation (1.1), and set up a computing procedure with various random decisions in it which more or less closely imitated the physical phenomenon under study.

This type of approach to distribution problems has long been known to statisticians under the name of "model sampling." The physicists have thought up a new name for it that seems likely to stick: the "Monte Carlo Method."

The formal solution of (1.1) is the truncated Neumann series

$$u_N = c + L(c) + L^2(c) + \cdots + L^{N-1}(c) + L^N(u_0), \qquad N > 0, \tag{1.2}$$

where L^K means the K-th iterate of L. Especial interest, of course, lies in the case in which the corresponding infinite series converges. If it does, it represents a function u which satisfies the equation

$$u = L(u) + c. \tag{1.3}$$

The error estimate is then provided by

$$u_N - u = L^N(u_0 - u). \tag{1.4}$$

Equation (1.2) shows that it makes no difference in the long run how u_0 was chosen, but Eq. (1.4) suggests that the nearer u_0 is to u, the faster the convergence will be.

The stochastic approach to estimating the solution of (1.1) will now be described in correspondingly general terms. Consider the space R over which the functions $c, u_0, u_1, \cdots$, are defined. Let x be a point in it at which it is desired

[1] Various practical aspects of the stochastic estimation are presented in [14].

to calculate u_N. A probability distribution is now set up on each of the spaces R, $R \times R$, $R \times R \times R$, $\cdots$, where $R \times R$ denotes the Cartesian product of R and R. Random variables Z_0, Z_1, Z_2, $\cdots$ are defined on these respective product spaces. The chain of probability distributions and of random variables is such that the sequence of conditional mean values[2] $v_0(x) = E(Z_0 \mid Z_0 = u_0(x))$ $v_1(x) = E(Z_1 \mid Z_0 = u_0(x))$, $v_2(x) = E(Z_2 \mid Z_0 = u_0(x))$, $\cdots$, satisfies (1.1). Since $v_0 = u_0$, it follows that $v_N(x) = u_N(x)$.

The computational problem then becomes one of calculating repeated realizations of Z_N and combining them into an appropriate statistical estimator of v_N.

The stochastic method is particularly well adapted to the case in which the value of $u_N(x)$ or $u(x)$ is to be estimated at only one point x.

In the case in which the interest lies in estimating the solution of (1.3), it is quite possible to carry the sequence Z_0, Z_1, Z_2, $\cdots$ on to infinity and define a random variable Z on the infinite product-space $R \times R \times \cdots$. But this has no significance for actual practice, and curiously enough, it turns out to be disadvantageous theoretically. That arises from the fact that mean values in statistics are ordinarily defined through absolutely convergent integrals and sums. This in turn imposes some irrelevant restrictions on L if v_∞ is to be identified with u, at least in the cases to be considered in this paper.[3] Therefore the attitude here in the "steady state" situation will be that a suitably large but finite N will be chosen once and for all and held fixed during the sampling.

The mathematical material preceding Section 7 is in the main a rearrangement and mathematical formulation of known procedures, presented so as to show up their relationships. The method of error analysis proposed in Section 4 and most of the material in Sections 7, 8, 9, and 10 are believed to be new. However, a good deal of work on the Monte Carlo method has been "published" in privately circulated, sometimes classified, reports, and one can never be quite sure of a priority under such circumstances.

A word of caution to the reader may be in order. The Monte Carlo method as a computational procedure has had its chief successes in problems which had natural stochastic bases and which were at the same time so complicated that they were inaccessible to ordinary analytic or numerical methods. This paper makes no pretense of putting the method into competition with the standard numerical practices, especially for the simpler type of problem for which good methods already exist. The idea here is merely to present some theory which may be of some interest for itself alone, and which unifies and clarifies certain of the Monte Carlo devices which have been proposed, and which lays the groundwork for further numerical experimentation aimed at investigating the limits of usefulness of the method for non-stochastic problems.

[2] We shall use the symbol $Pr(a \mid b)$ to denote the conditional probability of the event a, given that b has occurred, the symbol $E(Y)$ to denote the mean value of the random variable Y, and the symbol $E(Y \mid b)$ to denote the mean value of the conditional distribution of Y, given that b has occurred.

[3] As applied to the solution of simultaneous linear equations, the methods of the present paper were anticipated by Forsythe and Leibler [10] and Wasaw [16]. Both of these papers deal directly with the infinite product space. See also Curtiss [4].

We propose to stop short of describing the practical computational details. Except for the process of generating the necessary random elements, they can all be supplied easily by the experts in computing. The process of correctly generating the random elements and making sure that they are appropriately random could be the subject of another article at least as long as this one. If it were well done, it might contain substantial contributions to the philosophy of probability.

2. Specialization of the problem. The foregoing introduction has been very vague as to the nature of the operator L. Actually, the stochastic method to be presented here seems to be automatically confined to operators L of a Lebesgue-Stieltjes type,

$$L(u) = \int_R h(x, y)u(y)d_yk(x, y), \tag{2.1}$$

in which k is of bounded variation and h has integrability properties which will permit the iteration. We shall not pursue the question of generality any further, however. Instead we shall present the theory for two special cases: that in which L is an ordinary integral transform,[4]

$$L(u) = k\int_R h(x, y)u(y)\,dy, \tag{2.2}$$

and that in which L is a matrix and u is a vector.

The second case is the special case of (2.1) in which k is a step-function. Most of the exposition except for that in Section 4 will be directed toward this case. It is the more fundamental one in numerical analysis and the majority of the results can be carried over so readily to the continuous operator (2.2) that no comment on the matter will be necessary.[5]

It is convenient to introduce new notation for the matrix case. The space R will be thought of in this case as consisting of a finite discrete set of points $x_1, x_2, \cdots, x_n$. The function $u = u(x)$ will be represented by the vector

$$u = (u_1, u_2, \cdots, u_n) = (u(x_1), \cdots, u(x_n)).$$

Similarly, we write

$$u_N = (u_N(x_1), u_N(x_2), \cdots, u_N(x_n)) = (u_{N1}, u_{N2}, \cdots, u_{Nn}), \qquad N = 0, 1, 2, \cdots,$$

and

$$c = (c(x_1), c(x_2), \cdots, c(x_n)) = (c_1, c_2, \cdots, c_n).$$

Finally, the function $h(x, y)$ is represented by the matrix

$$H = [h(x_i, y_j)] = [h_{ij}].$$

Equation (1.1) becomes in matrix notation

$$u_{N+1} = Hu_N + c. \tag{2.3}$$

[4] The integral can be taken in either the Riemann or Lebesgue sense in the sequel.

[5] See Cutkosky [5].

The problem connected with this equation is to calculate one or more components of the vector u_N, given H, $u_0 = (u_{01}, u_{02}, \cdots, u_{0n})$, and c.

The series solution (1.2) becomes

$$u_N = c + Hc + H^2c + \cdots + H^{N-1}c + H^Nu_0 \tag{2.4}$$
$$= (I - H^N)(I - H)^{-1}c + H^Nu_0, N = 1, 2, \cdots.$$

Here the symbol I stands for the unit matrix. Of course the third member of the equation can be written down only if $I - H$ is non-singular.

It is well known that the necessary and sufficient condition for $\lim_{N\to\infty} H^N = 0$ is that all the eigenvalues of H (that is, the roots of the determinantal equation $|\lambda I - H| = 0$) must lie within the unit circle of the complex plane.[6] If this is the case—and we shall always assume that it is whenever we are discussing the situation as $N \to \infty$—then, $(I - H)^{-1}$ exists, and $u = \lim_{N\to\infty} u_N = (I - H)^{-1}c$, which satisfies the linear equations

$$u = Hu + c. \tag{2.5}$$

3. The solution of $Au = b$. We digress for a moment here to note the relationship between (2.3), (2.4), and (2.5), and the important problem of solving the system of linear equations

$$Au = b, \tag{3.1}$$

where $b = (b_1, \cdots, b_n)$ is an arbitrary vector. Choose the matrix H, and also a new one M, so that $H + MA = I$, and choose $c = Mb$. Then (2.5) reduces to $MAu = Mb$. If M is non-singular, then this system is precisely equivalent to (3.1) in the sense that each solution u of (2.5) is a solution of (3.1), and vice versa. If there is more than one solution to (3.1), then A is singular, and in this case it is easily checked that one of the eigenvalues of $H = I - MA$ is unity. This means that u_N in (2.3) and (2.4) cannot converge to the solution. We shall therefore exclude this case and assume that A is non-singular.

With M and A both non-singular, and with H having its eigenvalues all inside the unit circle, (2.3) becomes what is known in the theory of linear algebraic systems as a stationary linear iterative process, or Wittmeyer process, for solving $Au = b$. There are obviously an infinite number of ways of choosing M and H so that the conditions are fulfilled. One standard method is to split up A into the difference of two matrices V and W, where V is easily inverted. That is, let $A = V - W$. Then take $M = V^{-1}$, $H = V^{-1}W$. It is easily checked that $H + MA = I$, and it is not hard to arrange things so that the eigenvalues of H are all sufficiently small in modulus.[7]

[6] See [13, pp. 97–98].

[7] See [7, pp. 132–133]. The iterative method of Seidel takes V as a triangular matrix obtained by replacing all elements of A above the principal diagonal by zeros. The iterative method of Jacobi takes V as a principal diagonal matrix whose diagonal is that of A. The so-called relaxation method is *not* a linear process. So far as the author is aware, the extension of Monte Carlo methods to non-linear processes has not yet been accomplished, and may be impossible. For an interesting and scholarly classification of non-stochastic methods of solving linear equations, see Forsythe [9].

The calculation of A^{-1}, if that is the problem, can of course be accomplished by specializing b appropriately. Alternatively, the Wittmeyer process can be directly adapted to this problem by replacing c by M, and $u_0, u_1, u_2, \cdots$, by square matrices $U_0, U_1, U_2, \cdots$.

The importance of these remarks in the present context is merely that they show that with a little preliminary preparation any system $Au = b$ with a non-singular matrix can be solved by the Monte Carlo methods to be described hereinafter.

4. The stochastic methods for $N = 1$. The basic idea in the Monte Carlo attack on problems (1.1) and (1.3) will now be described for the case of the zero-th iteration of the recursion formula. To avoid dealing with a trivial problem, we present the ideas here with L taken as the continuous operator (2.2) rather than as a discrete operator.

The problem then is to estimate the numerical value of

$$u_1(x) = \int_R h(x, y)u_0(y)\, dy + c(x).$$

The function $c(x)$ will play no significant role in the present discussion, so we confine ourselves to the estimation of

$$I(x) = \int_R h(x, y)u(y)\, dy.$$

Now let functions $z(x, y)$ and $p(x, y)$ be chosen so that[8]

$$(1) \qquad zp = h,$$

$$(2) \qquad p \geqq 0,$$

$$(3) \qquad \int_R p(x, y)\, dx = 1.$$

Then for each x, p may be regarded as a probability density on R. Let X be a vector random variable with the probability distribution defined by p. Consider the random variable $Z = z(x, X)u(X)$. Clearly

$$E(Z) = \int_R z(x, y)u(y)p(x, y)\, dy$$

$$= \int_R h(x, y)u(y)\, dy = I(x).$$

As an example, let R be one dimensional, and let

$$h(x, y) = \begin{cases} e^{-xy}, & y \geqq 0 \\ 0, & y < 0 \end{cases}$$

Then $I(x) = \int_0^\infty e^{-xy}u(y)\, dy$, the Laplace transform of $u(y)$. A natural choice for $p(x, y)$ would be the Pearson Type III density function, e^{-xy}/x. Then $z(x, y) = x$, and $Z = xu(X)$.

[8] The stochastic mechanism which is being arranged here has been proposed by various writers. See Kahn [11], and also [14].

One of the standard statistical procedures for estimating the mean of the distribution of a random variable Z is to make ν independent observations on the random variable and then take their average $\bar{Z}$ as the estimator. If the standard deviation[9] σ of the distribution of the random variable is finite, then the probability distribution of $\bar{Z}$ is asymptotically Gaussian, or "normal", with (of course) the same average, and a standard deviation equal to $\sigma/\nu^{1/2}$. (This is a special case of the famous Central Limit Theorem; see Cramer [3, pp. 215–217]. The approximation is usually very close for $\nu \geqq 30$.)

More than 99 per cent of a normal distribution lies within the interval: [mean $\pm 3 \times$ standard deviation]. From this we can easily calculate the sample size ν theoretically necessary to achieve with this level of certainty a given statistical accuracy in using $\bar{Z}$ as an estimator of $I(x)$.

Let us say that we wish to be almost sure that $\bar{Z}$ will lie in the interval $I \pm \Delta I$. That is, we want to arrange things so that $Pr(|\bar{Z} - I| \leqq \Delta I) \geqq .99$. This means that $\Delta I \geqq 3\sigma/\nu^{1/2}$, from which we get

$$\nu \geqq \frac{9\sigma^2}{(\Delta I)^2} = \frac{9 \operatorname{Var}(Z)}{(\Delta I)^2}. \tag{4.1}$$

The formula shows that if ΔI is to be small, say 0.005, then unless Var (Z), is also small, ν will be well up in the hundreds of thousands. The two redeeming traits of the stochastic method are that (1) the sampling error and necessary sample size are independent of dimensionality and (2) they are independent of how locally smooth the integrand of $I(x)$ is. (It will be recalled that the error terms in the standard quadrature formulas involve high-order derivatives.)

But it is clearly worth while to consider methods of reducing the sample size. The statisticians have a number of devices for increasing the accuracy of sampling surveys, and almost all of them are applicable here. In the present paper we shall study only the procedures known as "sampling with probabilities in proportion to size," or "importance of sampling."[10] They take advantage of the fact that we have an infinite number of ways to choose z and p for any given problem, and an astute choice may decrease the variance by a surprising amount.

The variance of Z (and we henceforth assume that it is finite) is given by the formula

$$\sigma^2 = E(Z^2) - [E(z)]^2 = \int_R zhu^2 \, dy - I^2.$$

If the answer I were known in advance, and z were chosen to be I/u (p would then have to be hu/I, which would mean that the integrand hu was non-negative), then clearly $\sigma^2 = 0$. This leads us to propose the following arrangement as

[9] The standard deviation of Y is defined to be

$$\{E[(Y - E(Y))^2]\}^{1/2} = \{E(Y^2) - [E(Y)]^2\}^{1/2}.$$

The square of the standard deviation is called the variance; we shall write it as Var (Y).

[10] See Deming [6 pp. 92–93].

a guide in choosing z and p. First choose an integrable, non-negative function p' such that the function $\epsilon = \epsilon(x, y)$ defined by

$$hu - p' = p'\epsilon \tag{4.2}$$

is as small as possible in absolute value. The function p' is, moreover, to be chosen so that the integral $J = \int_R p'\, dy$ can be obtained numerically without too much trouble. Let $p = p'/J$ and let $z = h/p = J(1 + \epsilon)/u$. Then it is easily seen that z and p satisfy the conditions (1)–(3) listed above. The estimator Z is given by $Z = J(x)(1 + \epsilon(x, X))$.

The variance of $\alpha + \beta Y$, where α and β are constants and Y is a random variable, is simply β^2 Var (Y). Thus

$$\sigma^2 = \text{Var}\,(Z) = J^2 \text{ Var } [\epsilon(x, X)]. \tag{4.3}$$

If $e(x)$ is the least upper bound of $|\epsilon(x, y)|$ for Y on R, then it might be safe to presume that Var $(\epsilon) \leqq e^2/2$. (If the distribution of ϵ were rectangular with range $2e$—an unfavorable case—then Var $(\epsilon) = e^2/3$.) This appraisal gives us the formula

$$\nu \geqq \frac{4.5\, J^2 e^2}{(\Delta I)^2} \tag{4.4}$$

for the sample size theoretically required to achieve an accuracy of $\pm\Delta I$ with at least 99% certainty.[11]

It might be noted that (4.3) and (4.4) are quite independent of whether or not $|\epsilon|$ remains small for all y. If it does, then there is an implication that hu cannot go very far in the negative direction, since p' cannot be negative. These restrictions on the usefulness of the arrangement (4.2) can be circumvented to some extent, but we shall not go onto the matter here.

We have been carrying the somewhat superfluous parameter x throughout the above discussion mainly to emphasize the link between quadrature and other problems. The usual quadrature problem would of course be presented with $h(x, y) = h(y)$, $u(y) \equiv 1$. But it is perhaps worthwhile to observe that if there is a parameter x in the problem, then by choosing p so that it is dependent only on $x - y$, a set of determinations of X can be gotten once and for all, from which the statistics for Z can be computed over and over again for as many different values of x as may be desired.

One final remark which applies to all the remaining sections of this paper as well as to the present section is this: The statistical error was the only kind of error under consideration here, but of course there would be many other possible sources of error in the actual numerical applications. For example, there would be possible round-off errors, mistakes, and systematic errors of one kind or another. In accordance with the resolution expressed at the end of the Introduction, none of these non-statistical errors will be discussed.

[11] In any case, certainly Var $(\epsilon) \leqq e^2$.

5. The case $N > 1$. Suppose that the method of the preceding section had been applied to the trivial problem of estimating

$$u_1(x_i) = \sum_{j=1}^{n} h(x_i, x_j)u(x_j) = \sum_{j=1}^{n} h_{ij}u_j = Hu]_i,$$

for some fixed i, where the notation is that of Section 2.[12] The procedure would have been to select numbers $z_{ij} = z(x_i, x_j)$ and $p_{ij} = p(x_i, x_j)$ such that

$$(1) \quad z_{ij}p_{ij} = h_{ij}, \qquad i, j, = 1, \cdots, n,$$

$$(2) \quad p_{ij} \geqq 0, \qquad i, j, = 1, \cdots, n,$$

$$(3) \quad \sum_j p_{ij} = 1, \qquad i \quad = 1, \cdots, n;$$

then to let X be a vector random variable with the probability distribution given by $Pr(X = x_j) = p_{ij}$, and to form the random variable $Z = z(x_i, X)u(X)$ as before. Its mean value would be

$$E(Z) = \sum_j z(x_i, x_j)u(x_j)p(x_i, x_j) = u_{1i}.$$

We shall now show how to extend this stochastic scheme to the iterations of H.

We continue to define z_{ij} and p_{ij} as above. Consider a random walk on R defined as follows: the starting point X_0 is a random variable with a probability distribution given by $p_i = p(x_i)$, where $p_i > 0$, $\sum_i p_i = 1$, but p_i is otherwise arbitrary.[13] Thereafter the successive positions or states visited by the random walk are random variables $X_1, X_2, \cdots$, whose distributions are given by the formula $Pr(X_{k+1} = x_j \mid X_k = x_i) = p_{ij}$, $i, j = 1, \cdots, n$.

These directions have the effect of unambiguously specifying a probability distribution on the product space $\overbrace{R \times R \times \cdots \times R}^{N}$ for each N. The typical point is assigned the probability $p_i p_{ij_1} p_{j_1 j_2} p_{j_2 j_3} \cdots p_{j_{n-1} j}$. The chain of random variables $X_0, X_1, X_2, \cdots$ so defined is a simple Markov chain.[14]

Now define a new chain of random variables as follows:

$$Z_0 = Z_0[u] = u(X_0)$$

$$Z_1 = Z_1[u, z] = z(X_0, X_1)u(X_1)$$

$$Z_2 = Z_2[u, z] = z(X_0, X_1)z(X_1, X_2)u(X_2)$$

$$\vdots$$

$$Z_N = Z_N[u, z] = z(X_0, X_1) \cdots z(X_{N-1}, X_N)u(X_N).$$

$$\vdots$$

[12] By $b]_i$, where b is a vector, we mean the i-th component of b.

[13] This distribution does not play an intrinsic role in our discussion and is introduced only for logical completeness.

[14] See Feller [8, Chap. 15]. Actually our specification uniquely assigns a probability distribution to the infinite product space $R \times R \times \ldots$ (see [12, pp. 28–33]), but we shall not make use of this distribution.

We shall call $Z_0, Z_1, Z_2, \cdots$, an m-chain (m for multiplicative).
Let the vectors $v_0, v_1, \cdots$, b, defined as follows:

$$v_{0i} = v_0(x_i; u) = u(x_i),$$

$$v_{Ni} = v_N(x_i; u, z) = E(Z_N[u, z] \mid X_0 = x_i), \qquad N = 1, 2, \cdots$$

Consider now a typical path of the random walk represented by the sequence of random variables $X_0, X_1, \cdots, X_N$. Let this path be $x_i, x_{j_1}, x_{j_2}, \cdots, x_{j_{N-1}}, x_j$. The conditional probability that the random walk actually takes this path is $p_{ij_1}p_{j_1j_2}\cdots p_{j_{N-1}j}$. Therefore by the definition of the concept of mean value in the theory of probability,

$$v_{Ni} = \sum_{j_1=1}^{n}\sum_{j_2=1}^{n}\cdots\sum_{j_{N-1}=1}^{n}\sum_{j=1}^{n} z_{ij_1}z_{j_1j_2}\cdots z_{j_{N-1}j}u_j p_{ij_1}p_{j_1j_2}\cdots p_{j_{N-1}j}$$
$$= \sum_{j_1=1}^{n}\sum_{j_2=1}^{n}\cdots\sum_{j=1}^{n} h_{ij_1}h_{j_1j_2}\cdots h_{j_{N-1}j}u_j = H^N u]_i .$$

Thus

$$v_{N+1} = Hv_N, \qquad N = 0, 1, 2, \cdots, \tag{5.1a}$$

$$v_N = H^N u, \qquad N = 0, 1, 2, \cdots. \tag{5.1b}$$

The question of estimating u_N in the recursion relation (2.3) can now be resolved rather easily. It is clear that

$$u_{Ni} = (c + Hc + \cdots + H^{N-1}c + H^N u_0)]_i$$
$$= v_0(x_i, c) + v_1(x_i; c, z) + \cdots + v_{N-1}(x_i; c, z) + v_N(x_i; u_0, z)$$
$$= E(Z_0[c] + Z_1[c, z] + \cdots + Z_{N-1}[c, z] + Z_N[u_0, z] \mid X_0 = x_i),$$
$$= E(\Sigma_{N-1}[c, z] + Z_N[u_0, z] \mid X_0 = x_i),$$

where $\Sigma_N[u, z] = Z_0[u] + Z_1[u, z] + \cdots + Z_N[u, z]$. This shows that if we define vectors $w_0, w_1, w_2, \cdots$, by

$$w_{0i} = w_0(x_i; u) = u(x_i),$$

$$w_{Ni} = w_N(x_i; c, u, z) = E(\Sigma_{N-1}[c, z] + Z_N[u, z] \mid X_0 = x_i), \quad N = 1, 2, \cdots.$$

then with $u = u_0$, these vectors satisfy the relations

$$w_{N+1} = Hw_N + c, \qquad N = 0, 1, 2, \cdots, \tag{5.2a}$$

$$w_N = (I - H^N)(I - H)^{-1}c + H^N u_0, \quad N = 1, 2, \cdots, \tag{5.2b}$$

Furthermore, if we assume that the eigenvalues of H are in modulus all less than unity, then the vector $w_\infty = \mathrm{Lim}_{N\to\infty} w_N$ exists and is the solution of the system $u = Hu + c$.

Thus the conditional mean values of the chain of random variables $\Sigma_{N-1} + Z_N$, $N = 1, 2, \cdots$ provide a solution to our basic problem. The actual computation would consist in making a large number of realizations of the vector random variable $(X_0, X_1, \cdots, X_N)$, calculating $\Sigma_{N-1} + Z_N$ for each realization, and

averaging the results, or otherwise combining them into a statistical estimator of the mean value.

Several remarks are now in order.

(1) This method of solving linear equations was proposed by Wasow [16]. He considered only the infinite product space $R \times R \times \cdots$, and was thereby forced to use the restriction that the matrix of the absolute values of the elements of H had its eigenvalues in the unit circle. The pedigree of Wasow's suggestion goes back to a well-known paper of Courant, Friedrichs, and Lewy [1], in which the Green's function of an elliptic difference equation is identified with the mean number of visits to the points of the lattice made by a particle performing a certain random walk on the lattice. (See the remark (3) below.)

(2) Wasow and other writers describe the process of realizing Z_N as a weighted random walk in which the particle starts at x_i with a "mass" unity, which is then multiplied by the factor $z(x_i, x_{j_1})$, then multiplied by $z(x_{j_1}, x_{j_2})$, and so forth. He deals in [16] with matrix inversion and chooses $u_0 = 0$, so his $c = (\delta_{kj}, j = 1, 2, \cdots, n)$, where $\delta_{kj} = 1, j = k$, $\delta_{kj} = 0, j \neq k$. With this specialization, Σ_N then becomes in Wasow's words the total amount of mass carried through the point x_k during the N steps of the random walk.

(3) If in the function $\Sigma_N[c, z]$ we take $z_{ij} \equiv 1$, $c = (\delta_{kj})$, then Σ_N is the total number of visits to x_k made by the N-step random walk starting at x_i, counting in one visit for the starting point if $i = k$; and $w_{N+1}(x_i ; (\delta_{kj}), 0, 1)$ is the mean number of visits to x_k. (In making this observation we are of course no longer tying ourselves down to the requirement that $z_{ij}p_{ij} = h_{ij}$ where h_{ij} is given in advance, but instead we are assuming that the p_{ij} are given a priori.)

(4) The chain $Z_0, Z_1, Z_2, \cdots$ represents a type of branching process. It is a Markov chain which may be of any order from one to infinity, depending on the choice of the z's. Such chains have been studied by Montroll [15], together with the companion type obtained by replacing multiplications by additions. They have numerous applications to theoretical physics and physical chemistry.

(5) In the solution of $u = Hu + c$, or in the problem of inverting of $I - H$, it should be noted that once a large set of realizations of $(X_0, X_1, X_2, \cdots, X_N)$ has been obtained with all the random walks starting from some fixed x_i, then by proper bookkeeping procedure they can be used a number of times. Not only can they be used to get the statistics for a number of different vectors c—say, all the columns of the identity matrix I, which would be the procedure for inverting $I - H$—but also, by considering a visit to $x_j, j \neq i$, as starting a new random walk, they can be used to find components of $u_{N'i}, N' < N$, or of u, other than the i-th component. Nevertheless a peculiarity of the stochastic method here presented is that it seems to appear to the best advantage in comparison with the standard deterministic methods when the problem is to find only one component of u_N or of u, or one row of $(I - H)^{-1}$. This fact has already been commented upon once before, in the Introduction.

(6) A final remark of minor importance is that the estimator used in the quadrature problem of Section 4 was analogous to the Z_1 in the present section, not to Z_0. Thus to achieve strict parallelism between the two sections, we should

have assigned an appropriate distribution to $R \times R$ in Section 4 rather than just to R, and then considered a conditional distribution in $R \times R$ under a hypothesis relating to R.

6. The m-chain method. The development of the preceding section can easily be modified so that an m-chain does the work of the random sequence $\Sigma_0 + Z_1, \Sigma_1 + Z_2, \cdots$. To accomplish this, we adjoin n additional points $x_{n+1}, x_{n+2}, \cdots, x_{2n}$ to R, and set up the $2n \times 2n$ partitioned matrix

$$H^* = \begin{bmatrix} h^*(x_i, x_j) \\ j = 1, \cdots, 2n \\ j = 1, \cdots, 2n \end{bmatrix} = \left[\begin{array}{c|c} H & I \\ \hline 0 & I \end{array}\right]$$

Also, we let $(u_0 : c) = (u_{01}, u_{02}, \cdots, u_{0n}, c_1, c_2, \cdots, c_n)$, with $c_i = c(x_i) = c(x_{i+n})$, $i = 1, \cdots, n$. It is to be noticed that

$$H^{*N} = \left[\begin{array}{c|c} H^N & I + H + H^2 + \cdots + H^{N-1} \\ & = (I - H^N)(I - H)^{-1} \\ \hline 0 & I \end{array}\right]. \tag{6.1}$$

Now let z^*_{ij} and p^*_{ij} be chosen so that $z^*_{ij} p^*_{ij} = h^*_{ij}$, $p^*_{ij} \geqq 0$, $\sum_{j=1}^{2n} p^*_{ij} = 1$. Set up the random walk $X_0^*, X_1^*, \cdots$, on the lattice $x_1, \cdots, x_n, x_{n+1}, \cdots, x_{2n}$, using the transition probabilities p^*_{ij}. Then construct the new m-chain

$$Z_0^* = Z_0^*[(u:c)] = \begin{cases} u(X_0^*), & X_0^* = x_1, \cdots, x_n, \\ c(X_0^*), & X_0^* = x_{n+1}, \cdots, x_{2n}. \end{cases}$$

$$Z_1^* = Z_1^*[(u:c), z^*] = z^*(X_0^*, X_1^*) \times \begin{cases} u(X_1^*), & X_1^* = x_1, \cdots, x_n, \\ c(X_1^*), & X_1^* = x_{n+1}, \cdots, x_{2n} \end{cases}$$

$$\vdots$$

$$Z_N^* = Z_N^*[(u:c), z^*]$$

$$= z^*(X_0^*, X_1^*) \cdots z^*(X_{N-1}, X_N^*) \times \begin{cases} u(X_N^*), & X_N^* = x_1, \cdots, x_n, \\ c(X_N^*), & X_N^* = x_{n+1}, \cdots, x_{2n}. \end{cases}$$

Let the vectors $v_0^*, v_1^*, \cdots$, be defined as follows:

$$v^*_{0i} = v_0^*(x_i; (u:c)) = \begin{cases} u_i, i = 1, 2, \cdots, n, \\ c_{i-n}, i = n+1, n+2, \cdots, 2n. \end{cases}$$

$$v^*_{Ni} = v_N^*(x_i; (u:c), z^*) = E(Z_N^*[(u:c), z^*] \mid X_0^* = x_i), N = 1, 2, \cdots$$

Then according to (5.1) these vectors satisfy the relations

$$v^*_{N+1} = H^* v^*_N, \qquad N = 0, 1, 2, \cdots, \tag{6.2a}$$

$$v^*_N = H^{*N}(u:c), \qquad N = 0, 1, 2, \cdots. \tag{6.2b}$$

The components of these vectors for $i = n+1, \cdots, 2n$ deserve some attention. From the definition of v_0^*, $v^*_{0,n+i} = c_i$, $i = 1, \cdots, n$. As for v_1^*, it will be

seen by looking at H^* that $z_{ij}^* p_{ij}^* = \delta_{ij}$, $i > n, j > n$. This means that $v_{1,n+i}^* = v_{2,n+i}^* = \cdots = v_{N,n+i}^* = c_i$, $i = 1, \cdots, n$.

Using (6.1) to (6.2) and these facts, we obtain, letting $u = u_0$

$$v_{0i}^* = \begin{cases} u_{0i} & i = 1, \cdots, n, \\ c_{i-n} & i = n+1, n+2, \cdots, 2n, \end{cases} \tag{6.3a}$$

$$v_{N+1,i}^* = \left.\begin{cases} [Hv_N^*]_i + c_i, & i = 1, \cdots, n \\ c_{i-n}, & i = n+1, \cdots, 2n \end{cases}\right\} N = 0, 1, \cdots \tag{6.3b}$$

$$v_{Ni}^* = \left.\begin{cases} [H^N u_0]_i + (I - H^N)(I - H)^{-1} c]_i & i = 1, \cdots, n \\ c_{i-n}, & i = n+1, \cdots, 2n. \end{cases}\right\} N = 1, 2, \cdots \tag{6.3c}$$

(In forming the product of the non-conformable factors H and v_N^*, only the first n components of v_N^* are to be used.)

These relations show that the first n components of v_N^* satisfy the recursion relation (2.3) with the initial vector u_0. If $H^N \to 0$, then $v_\infty^* = \operatorname{Lim}_{N\to\infty} v_N^*$ exists and its first n components satisfy the linear equations (2.5).

Although the vectors w_N of the preceding section and v_N^* of this section must be identical, they are based on quite different stochastic processes. For one thing, in the previous section it was understood that $\sum_{j=1}^{n} p_{ij}$ always equaled unity, which implied that the random walk remained forever on $x_1, \cdots, x_n$, whereas now the corresponding sum of probabilities must be less than unity, and the random walk will not necessarily remain forever on the first n points of the lattice.

In fact, if we choose $p_{ij}^* = \delta_{ij}$ for $i > n$, then the new points x_{n+1}, $x_{n+2}, \cdots, x_{2n}$ become trap states or absorbing states for the random walk X_0^*, $X_1^*, \cdots$. It is now a random walk with absorbing barriers.

To increase the comparability of the processes of this section and of the preceding one for solving (2.3) and (2.5) we can set up the solution of the preceding section with $H^{*\prime}$ replacing H, z^* replacing z, p^* replacing p, $(u_0 : 0) = (u_{01}, u_{02}, \cdots, u_{0n}, 0, \cdots, 0)$ replacing u_0, and

$$(c : 0) = (c_1, c_2, \cdots, c_n, 0, \cdots, 0)$$

replacing c. The vectors $(u_0 : 0)$ and $(c : 0)$ have the effect of annulling the right-hand blocks of H^{*N}. Then the process $\Sigma_{N-1}^*[(c : 0), z^*] + Z_N^*[(u_0 : 0), z^*]$, where $\Sigma_N^* = Z_0^* + \cdots + Z_N^*$, will have the same conditional mean values for $X_0^* = x_i$, $i \leqq n$, as $Z_N^*[(u_0 : c), z^*]$. The random variable $\Sigma_{N-1}^* + Z_N^*$ is identically equal to zero if the random walk ever visits the new part of the lattice (that is, the point-set $x_{n+1}, x_{n+2}, \cdots, x_{2n}$) in the course of the first N steps.

The m-chain method of inverting matrices was first proposed by Forsythe

and Leibler [10], following a suggestion of von Neumann. They considered only the infinite product space $R \times R \times \cdots$ and thus had to make a restriction on H similar to that which Wasow used.

7. Variances. The variance of the random variable $Z_N[u, z]$ of section 5 is easily derived. Let $K = [z_{ij}\ h_{ij}] = [h_{ij}^2/p_{ij}]$. Then by the argument which led to our fundamental equations (5.1), we find that

$$(7.1) \qquad {}_2v_{Ni} = {}_2v_N(x_i\,;\,u, z) = E\{(Z_N[u, z])^2 \mid X_0 = x_i\} = K^N u^2]_i\,,$$

where u^2 means the vector $(u_1^2, u_2^2, \cdots, u_n^2)$. (We shall frequently use this exponential notation to represent the operation of squaring the components of a vector.)

The conditional variance of Z_N, given that $X_0 = x_i$, is of course the vector ${}_2v_N - v_N^2$.

Turning to the random variable $\Sigma_{N-1}[c, z] + Z_N[u, z]$, we first find by using the method that led to (5.1) that for $s \leqq t$,

$$E(Z_s[c, z]Z_t[u, z] \mid X_0 = x_i) = K^s C H^{t-s} u]_i\,,$$

where C is a principal diagonal matrix whose diagonal elements are c_1, $c_2, \cdots, c_n$. We apply this relation and (7.1) to the expansion of the square in the last member of

$$\begin{aligned} {}_2w_{Ni} &= {}_2w_N(x_i\,;\,c, u, z) = E\{(\Sigma_{N-1}[c, z] + Z_N[u, z])^2 \mid X_0 = x_i\} \\ &= E\{(Z_0[c] + Z_1[c, z] + \cdots + Z_{N-1}[c, z] + Z_N[u, z])^2 \mid X_0 = x_i\}. \end{aligned}$$

By going through some matrix algebra, the following rather formidable (but equivalent) formulas can be reached:

$$\begin{aligned} (7.2a) \qquad {}_2w_N &= (I - K)^{-1}c^2 + K^N[u^2 - (I - K)^{-1}c^2] \\ &\quad + 2(I - K)^{-1}[C(I - H^{N-1})(I - H)^{-1}Hc - S_{N-1}c] + 2S_{N-1}Hu, \end{aligned}$$

$$\begin{aligned} (7.2b) \qquad {}_2w_N &= K^{N-1}c^2 + K^N u^2 + (I - K)^{-1}(I - K^{N-1})[2C(I - H)^{-1}c - c^2] \\ &\quad - (CH^{N-1} + S_{N-2})(I - H)^{-1}Hc + 2S_{N-1}Hu, \end{aligned}$$

where $S_N = K^N CH + K^{N-1}CH^2 + \cdots + K^2 CH^{N-1} + KCH^N$.

The variance is of course to be obtained by subtracting the vector w_N^2 from the vector ${}_2w_N$.

The limiting forms of these formulas will now be derived, under the hypothesis that the eigenvalues of both K and H lie inside the unit circle in the complex plane. For this we need the following simple result.

Lemma. If A is any $n \times n$ matrix with complex elements, and with the property that all of its eigenvalues lie in the unit circle, then there exist constants $m = m(A) > 0$ and $r = r(A) > 1$ which are independent of N, and are such that $|\, a_{ij}^{(N)} \,| < m/r^N$, $i, j = 1, 2, \cdots, n$, $N = 1, 2, \cdots$ where $a_{ij}^{(N)}$ is the element in the i-th row and the j-th column of A^N.

For the proof, we first observe that the eigenvalues of rA, where r is now any scalar, $r \neq 0$, are those of A multiplied by r, because if y is an eigenvector for the eigenvalue λ, then the equations $Ay = y$ and $rAy = r\lambda y$ are equivalent. This means that if A satisfies the hypothesis of the Lemma, then a real number r exists, $r > 1$, such that the eigenvalues of rA are still less than unity in modulus. Therefore $\mathrm{Lim}_{N\to\infty} (rA)^N = 0$. Therefore there is a positive number m, independent of N, such that for all N, $| r^N a_{ij}^{(N)} | < m$. The result follows at once from this.

We now apply the Lemma to S_N. Let $\bar{c} = \max_i | c_i |$. Then[15]

$$| (K^s C)_{ij} | < m(K)\bar{c}/[r(K)]^s,$$

and

$$| (K^s C H^{N-s+1})_{ij} | < nm(K)m(H)\bar{c}/[r(K)]^s[r(H)]^{N-s+1}, 0 \leqq s \leqq N.$$

With $\bar{m} = \max [m(K), m(H)]$, $\bar{r} = \min [r(K), r(H)]$, this becomes

$$| (K^s C H^{N-s+1})_{ij} | < n\bar{m}^2\bar{c}/\bar{r}^{N+1}.$$

From this we get the fact that *if the eigenvalues of K and H are all less than unity in modulus, then there exist numbers $\bar{m} = \bar{m}(K, H, C) > 0$ and $\bar{r} = \bar{r}(K, H) > 1$ independent of N such that*

$$| (S_N)_{ij} | < nN\bar{m}/\bar{r}^N; \qquad i, j = 1, \cdots, m; \qquad N = 1, 2, \cdots. \tag{7.3}$$

(This estimate could be refined so that it is independent of n, using for example the methods described in [2, p. 16 (footnote)], but it is sufficiently precise as it stands for present purposes.)

The inequality (7.3) implies, of course, that $\mathrm{Lim}_{N\to\infty} S_N = 0$.

Therefore the limit of the vector ${}_2w_N$, which we shall denote by ${}_2w_\infty$, is given by the formulas

$${}_2w_\infty = (I - K)^{-1}c^2 + 2(I - K)^{-1}C(I - H)^{-1}Hc, \tag{7.4a}$$

$${}_2w_\infty = (I - K)^{-1}[2C(I - H)^{-1}c - c^2]. \tag{7.4b}$$

It is of interest to specialize the m-chain variance given through (7.1) to the augmented $2n \times 2n$ matrix H^* discussed in Section 6. Let z_{ij}^* and p_{ij}^* be chosen as previously, but with the proviso now that $p_{ij}^* = 0$ whenever $h_{ij}^* = 0$, $i = 1, 2, \cdots, n$, $j = n + 1, n + 2, \cdots, 2n$, and also that $p_{ij}^* = \delta_{ij}$, $i > n$, $j > n$. Then the matrix $K^* = [z_{ij}^* h_{ij}^*]$ has the appearance

$$K^* = \left[\begin{array}{c|ccc} z_{ij}^* h_{ij} & 1/p_1^* & & 0 \\ i = 1, \cdots n & & 1/p_2^* & \\ j = 1, \cdots n & 0 & & 1/p_N^* \\ \hline 0 & & I & \end{array}\right]$$

[15] We are using the notation B_{ij} for the i,j-th element of the matrix B.

where $p_i^* = 1 - \sum_{j=1}^{n} p_{ij}^*$. It follows as in the derivation of (6.2) that if we define

$$ {}_2v_{0i}^* = \begin{cases} u_i^2, & i = 1, \cdots, n, \\ c_{i-n}^2, & i = n+1, \cdots, 2n. \end{cases} $$

$$ {}_2v_{Ni}^* = E\{(Z_N^*[(u:c), z^*])^2 \mid X_0^* = x_i\}, \qquad N = 1, 2, \cdots, $$

then letting $u = u_0$,

$$ (7.5a) \qquad {}_2v_{N+1,i}^* = \begin{cases} K({}_2v_N^*)]_i + Qc^2]_i, & i = 1, \cdots, n \\ c_{i-n}^2, & i = n+1, \cdots, 2n \end{cases} \Bigg\} N = 0, 1, \cdots, $$

$$ (7.5b) \qquad {}_2v_{N,i}^* = \begin{cases} [K^N u_0^2]_i + (I - K^N)(I - K)^{-1} Qc^2]_i \\ c_{i-n}^2, \end{cases} $$

$$ \left.\begin{matrix} i = 1, \cdots, n \\ i = n+1, \cdots, 2n \end{matrix}\right\} N = 1, 2, \cdots. $$

Here Q denotes the principal diagonal matrix in the upper right-hand corner of K^*, and K means the $n \times n$ matrix[16] $[z_{ij}^* h_{ij}]$.

The limiting form of (7.5) is

$$ (7.6a) \qquad {}_2v_{\infty i}^* = K({}_2v_\infty^*)]_i + Qc_i^2, \qquad i = 1, \cdots, n $$

$$ (7.6b) \qquad {}_2v_{\infty i}^* = (I - K)^{-1} Qc^2]_i, \qquad i = 1, \cdots, n $$

It was noted in section 6 that the process

$$ \Sigma_{N-1}^*[(c:0), z^*] + Z_N^*[(u_0:0), z^*] $$

has the same conditional mean value as $Z_N^*[(u:c), z^*]$ for $X_0^* = x_i$, $i \leqq n$. With K redefined as above to mean $z_{ij}^* h_{ij}$, (7.2a) (7.2b), (7.4a), and (7.4b) give the components for $i \leqq n$ of the second moment of the random variable $\Sigma_{N-1}^* + Z_N^*$ exactly as they stand now. The components with $i > n$ are all zero.

Several remarks will now be made.

(1) The formula (7.6b) was derived by Forsythe and Leibler [16] as the second moment of the conditional distribution of the random variable Z_∞^* defined on the infinite product-space $R \times R \times \cdots$.

(2) The formulas for the second moments suggest that to keep down the variances, the numbers z_{ij} and p_{ij} or z_{ij}^* and p_{ij}^* should be chosen so that on the average (speaking intuitively) the elements of the matrix K should be as small in absolute value as possible. One way *not* to achieve this end is by letting p_{ij} or p_{ij}^* be positive when $h_{ij} = 0$ or $h_{ij}^* = 0$, because the unnecessary positive values of the p's in a given row could be apportioned out to the other elements of K or K^* in that row so as to make the elements h_{ij}^2/p_{ij} or h_{ij}^{*2}/p_{ij}^* a little

[16] To use K in this way is slightly inconsistent with the previous definition of K; perhaps K^{**} should be used.

smaller. (We have anticipated this remark to some extent by choosing $p_{ij} = 0$ when $h_{ij} = 0$ for $j > n$.)

(3) A corollary of our results of some slight interest for matrix theory is this. *If $[h_{ij}]$ is a real $n \times m$ matrix with all of its eigenvalues inside the unit circle, and if there exists a set of positive numbers p_{ij}, $i, j = 1, \cdots, n$ such that $\sum_j p_{ij} = 1$, $i = 1, \cdots, n$ and the eigenvalues of $[h_{ij}^2/p_{ij}]$ are all inside the unit circle, then the diagonal elements of $(I - H)^{-1}$ are all greater than $\frac{1}{2}$.* This follows from (7.4b) with c specialized to $c = (\delta_{kj}, j = 1, \cdots, n)$, $k = 1, \cdots, n$.

(4) The natural statistical estimators of the mean values v_N, w_N, v_N^*, are of course the arithmetic averages of many determinations of the respective random variables of which these are the theoretical mean values. Clearly the variances of these estimators always exist for finite matrices and finite values of N, no matter where the eigenvalues of H or H^* lie. Therefore in particular, the formula (4.1) for the number of random walks which must be performed to attain a given theoretical accuracy is here valid, with ΔI replaced by Δv_N, Δw_N, etc., as the case may be.

8. Comparison of the two methods of inverting matrices. We shall now use the formulas of the preceding section to effect a comparison of the statistical error of the m-chain method of inverting matrices (Forsythe-Leibler) to that of the method based on $\Sigma_{N-1} + Z_N$ (Wasow) given in Section 5.

It will be impossible to compare these methods unless the z's and p's are chosen comparably. Therefore for both methods we shall suppose that the arrangement described in Section 6 has been set up; that is, the one which uses H^*, z_{ij}^*, p_{ij}^*, X_N^*, and Z_N^*. The comparison will be made by comparing only the limits of the variances as $N \to \infty$, as the finite case seems to be rather intractible.

With $K = [z_{ij}^* h_{ij}, i, j = 1, \cdots, n]$, and Q defined as for (7.5), the formulas to compare are

$$ {}_2w_\infty = (I - K)^{-1}[2C(I - H)^{-1}c - c^2] $$

and

$$ {}_2w_\infty^* = (I - K)^{-1}Qc^2. $$

The specialization to the case of matrix inversion is accomplished by letting $c = (\delta_{kj}, j = 1, \cdots, n)$; then if the random walks start from $x = x_i$, we shall be estimating $(I - H)_{ik}^{-1}$. We then have

$$ (8.1) \qquad {}_2w_{\infty i} = (I - K)_{ik}^{-1}[2(I - H)_{ik}^{-1} - 1], \qquad {}_2w_{\infty i}^* = (I - K)_{ik}^{-1}/p_k^* . $$

The ratio of these second moments is

$$ (8.2) \qquad \frac{{}_2w_{\infty i}}{{}_2w_{\infty i}^*} = p_k^*[2(I - H)_{ik}^{-1} - 1] $$

Since $(I - H)_{ik}^{-1}$ is fixed by the terms of the problem, we have at our disposal only the numbers p_k^* in this formula. It is clear that for a given H these can always be chosen so that the m-chain method (Forsythe-Leibler) is poorer than

the other method (Wasow). And for an H such that the diagonal elements of $(I - H)^{-1}$ are small, the Wasow method might have a smaller variance than the other method, no matter how the random walk was set up.

These negative observations come far from telling the whole story. If the p_k^*'s are adjusted so as to be quite large, in an attempt to make the m-chain method show up favorably in (8.2), then the numbers p_{ij}^*, $i \leqq n, j \leqq n$, will be small and the factors z_{ij}^* will have to be proportionately larger. This will make the elements of K^{*N} and $(I - K)^{-1}$ large, and the values of both ${}_2w_\infty$ and ${}_2v_\infty$ will be large, which of course is undesirable. On the other hand, if the p_{ij}^*'s are chosen so that the values of p_k^* are small, then the mean duration of the random walk becomes very long (see Section 9 below), and this is perhaps more undesirable for the Wasow method than for the m-chain method because the former requires a little more computing per step in the random walk.

The question of an optimum method from the point of view of minimizing computing has not as yet been settled. Probably there is no such thing, because as in the case $N = 1$ treated in Section 4, it will turn out that the more one knows about the solution, the better one can do. This phenomenon will again be in evidence in Section 10.

9. The duration of the random walk. We shall now apply the formulas of the preceding section to obtain theoretical mean-values and dispersions in two classical problems connected with random walks, which have a bearing on the usefulness of the Monte Carlo methods discussed in this paper.

We shall be considering the random walk $X_0^*, X_1^*, \cdots$, on the lattice $x_1, \cdots, x_n, x_{n+1}, \cdots, x_{2n}$, Section 6. However, in considering the associated matrices $H^* = [z_{ij}^* p_{ij}^*]$ and $K^* = [z_{ij}^{*2} p_{ij}^*]$, (the latter was introduced in Section 7), the point of view will be that the transition probabilities p_{ij}^* and the factors z_{ij}^* are chosen first, and the values of h_{ij}^* are determined thereby, instead of the other way around. We suppose that $p_{ij}^* = \delta_{ij}$, $i > n, j > n$.

In each of the problems to be considered, $z_{ij} \equiv 1$. This means that $K^* = H^* = [p_{ij}^*]$, $K = H = [p_{ij}^*, i, j = 1, \cdots, n]$. The first of these matrices is a stochastic matrix. We denote the second one by P.

The first problem is that of determining the mean value and variance of the number of visits to x_k if the random walk starts at x_i. In accordance with remark (3), in Section 5, the total number of visits to x_k in an N-step random walk starting at x_i is

$$\Sigma_N^*[(\delta_{kj}), 1] = \delta_k(X_0^*) + \delta_k(X_1^*) + \delta_k(X_2^*) + \cdots + \delta_k(X_N^*),$$

where $\delta_k(x_j) = \delta_{kj}$, $j = 1, 2, \cdots, 2n$. The answers to the problem are thus given by (5.2) and (7.2a) or (7.2b).

Let the mean number of visits to x_k if the random walk started at x_i be denoted by $G_{ik}^{(N)}$. Then

$$\begin{aligned} G_{ik}^{(N)} &= E(\Sigma_N^*[(\delta_{kj}, 1] \mid X_0^* = x_i) \\ &= [(I - P^{N+1})(I - P)^{-1}]_{ik}, \qquad i \leqq n, k \leqq n, \\ G_{ik}^{(N)} &= \delta_{ki}, \qquad k > n. \end{aligned}$$

Assuming that $\lim_{N\to\infty} P^N = 0$, we have

$$G_{ik}^{(\infty)} = \lim_{N\to\infty} G_{ik}^{(N)} = (I - P)_{ik}^{-1}, \qquad i \leqq n, k \leqq n, \tag{9.1a}$$

$$G_{ik}^{(\infty)} = \delta_{ki}, \qquad k > n. \tag{9.1b}$$

Then by substituting into (7.4b), we obtain[17]

$$\lim_{N\to\infty} \operatorname{Var}(\Sigma_N^*[(\delta_{kj}), 1] \mid X_0^* = x_i) = 2G_{ik}^{(\infty)}G_{kk}^{(\infty)} - G_{ik}^{(\infty)} - (G_{ik}^{(\infty)})^2.$$

The first two terms of the formula are of course just a special case of (8.1).

The other problem concerns the duration of the random walk $X_0^*, X_1^*, \cdots$. We define the duration to mean the total number of visits made to each of the transient states $x_1, x_2, \cdots, x_n$ before absorption takes place. In counting the visits, we count in the starting point of the walk as one visit.

If the walk is limited to $N + 1$ visits (that is, to N steps), then the duration is clearly

$$\Sigma_N^*[(1;0), 1], \quad \text{where} \quad (1;0) = (\underbrace{1, 1, \cdots, 1}_{n}, \underbrace{0, 0, \cdots, 0}_{n}).$$

Its conditional mean value for $X_0^* = x_i$ is the vector d_N whose i-th component is

$$d_{Ni} = \sum_{j=1}^n G_{ij}^{(N)} = \sum_{j=1}^n [(I - P^{N+1})(I - P)^{-1}]_{ij}, \qquad i \leqq n,$$

$$d_{Ni} = 1, \qquad i > n.$$

The vector d_N satisfies the recursion relation

$$d_{N+1,i} = Pd_N]_i + 1, \qquad i \leqq n, N = 0, 1, 2, \cdots,$$

with $d_{0i} = 1$. Its limiting value is

$$d_{\infty i} = \sum_{j=1}^n (I - P)_{ij}^{-1}, \qquad i \leqq n. \tag{9.2}$$

The variance of the duration can be obtained by substituting appropriately into (7.2a) or (7.2b). We shall again write out only the limiting case. This is obtained easily from (7.4b); the i-th component is

$$\begin{aligned} {}_2d_{\infty i} - d_{\infty i}^{2'} &= 2\sum_{j=1}^n G_{ij}^{(\infty)} d_{\infty j} - d_{\infty i} - d_{\infty i}^2 \\ &\leqq d_{\infty i}(2d - 1 - d_{\infty i}), \end{aligned}$$

where $d = \max_i d_{\infty i}$.

The duration of non-truncated random walks in an infinite product space has recently been investigated in very general cases by Wasow [17]. Previously, special cases had been studied at some length by statistical theorists in connection with the problem of the mean length of a sequential test. (See [4] for further results and references.)

The conditional probability that X_N^* falls on one of the states $x_1, \cdots, x_n$,

[17] By $\operatorname{Var}(Y \mid b)$ we mean the variance of the conditional distribution of the random variable Y, given that the event b has occurred.

given that $X_0^* = x_i$, is clearly $\sum_{j=1}^{n}(P^N)_{ij}$. This is the probability that the walk lasts for N steps without falling into a trap state. To assert that $\lim_{N\to\infty} P^N = 0$ is therefore equivalent to saying that the probability that the walk lasts more than N steps approaches zero with N. In fact, if we consider the infinite product-space $R \times R \times \ldots$ for a moment, $\lim_{N\to\infty} P^N = 0$ means that the walk is "almost certainly" of finite duration, and conversely. It is known (see for example Curtiss [4, Section 11]), that if from each one of the states $x_1, x_2, \cdots, x_n$ it is possible to reach a trap state over a path in $R \times R \times \cdots$ with non-zero probability, then the walk is almost certainly of finite duration.

This provides a sufficient condition for the validity of (9.1) and (9.2). If it is satisfied, then all the eigenvalues of P will be less than unity in absolute value.

We mentioned in Section 8 that if the numbers $p_k^* = 1 - \sum_{j=1}^{n} p_{kj}^*$ were small, then the mean duration would tend to be large. This follows intuitively from the fact that the smaller the absorption probabilities are, the longer the walk will go on. A somewhat more rigorous demonstration can be given by considering the dependence of the eigenvalues of P on the row-sums of P. If $p_{ij}^* > 0$, $i = 1, \cdots, n$, $j = 1, \cdots, n$, then as the minimum row-sum approaches unity, the eigenvalue of maximum absolute value[18] approaches unity, and thus the convergence to zero of P^N becomes slower and slower.

10. Importance sampling. In this section we shall discuss the problem of the control of the statistical error. The treatment will be analogous to the one in the latter part of Section 4. It will pertain only to the estimators Z_N and Z_N^*, and not to the estimator $\Sigma_{N-1} + Z_N$. We shall assume throughout that the problem is to estimate the solution of (2.3) or (2.5). Therefore the factors z_{ij} and p_{ij}, or z_{ij}^* and p_{ij}^*, will always be related in the usual way to the elements of the matrix H or H^*.

It is worth while first to inquire into the conditions under which an m-chain can be a zero-variance estimator. It will suffice here to examine the situation only for the function $Z_N[u, z]$ of Section 5.

If Z_N did indeed have a conditional variance of zero, given that the random walk starts at x_i, then the value assumed on every path with positive probability would depend only on x_i. One way to insure this would be to choose $z_{ij} = \lambda u_i/u_j$, where λ is a constant; then

$$Z_N = \lambda^N \frac{u(X_0)}{u(X_1)} \cdot \frac{u(X_1)}{u(X_2)} \cdots \frac{u(X_{N-1})}{u(X_N)} u(X_N) = \lambda^N u(X_0).$$

Conversely, it can be shown that this is roughly the most general choice of the factors z_{ij} which will insure the desired result. We shall not try to give an accurate formulation of the theorem here.

Now if $z_{ij} = \lambda u_i/u_j$, then the requirement that $z_{ij}p_{ij} = h_{ij}$ implies that the choice of p_{ij} must be $h_{ij}u_j/u_i$. This in turn requires not only that all of these quotients must be real and non-negative, but also that $\sum_j h_{ij}u_j/\lambda u_i = 1$. The

[18] It happens to be real and positive, since P is a matrix of positive elements.

latter relation states that u must be an eigenvector of H corresponding to the eigenvalue λ.

If we stop insisting on stationary transition probabilities for our random walk, and also permit the factors z_{ij} to vary from step to step, then a zero-variance m-chain estimator of $u_N = H^N u_0$ can very easily be constructed for any u_0, provided that $h_{ij} > 0$ and $u_j > 0$, $i, j = 1, \cdots, n$. We simply choose

$$p_{ij}^{(1)} = \frac{h_{ij}\, u_{N-1,j}}{u_{N,i}}, \qquad z_{ij}^{(1)} = \frac{u_{N,i}}{u_{N-1,j}},$$

$$p_{ij}^{(2)} = \frac{h_{ij}\, u_{N-2,j}}{u_{N-1,i}}, \qquad z_{ij}^{(2)} = \frac{u_{N-1,i}}{u_{N-2,j}},$$

$$\vdots \qquad\qquad \vdots$$

$$p_{ij}^{(N)} = \frac{h_{ij}\, u_{0j}}{u_{1i}}, \qquad z_{ij}^{(N)} = \frac{u_{1i}}{u_{0j}}.$$

It is easily checked that Z_N formed with these factors $z_{ij}^{(K)}$ and with $u = u_0$, has the constant value $u_{N,i}$ for any random walk starting at x_i. For each K, $v_{K+1} = Hv_K$, where $v_{K,i} = E(Z_K \mid X_0 = x_i)$, but the chain is not a zero-variance chain unless $K = N$.

The argument can be extended so as to allow zero elements in H and u_0 by giving a little attention to the undefined quotients.

It is not difficult to proceed from here to a practical arrangement whereby given only approximate values of $u_0, u_1, u_2, \cdots, u_N$, a chain can be set up in which the variance can be exhibited as a function of the errors of the approximations, as was done in the case $N = 1$ in Section 4. Because of limitations of space we shall not pursue the matter further. Instead we shall study a special class of m-chains which use stationary transition probabilities and for which the statistical error analysis is usually easy to make. These m-chains, however, have the slight disadvantage that they are connected with a special type of matrix H.

The main problem is, as usual, to estimate the vector u_N in the recursion relation

$$u_{N+1} = Hu_N + c. \tag{10.1}$$

We shall also be interested in the problem of estimating u in the equation

$$u = Hu + c, \tag{10.2}$$

but as elsewhere in this paper this problem will be considered as the limiting case of the first problem as N becomes infinite. Whenever (10.2) is in view, we shall as usual assume that all the eigenvalues of H lie inside the unit circle.

We now impose certain conditions on H. They are that $h_{ij} \geqq 0$, and that a vector u_0 with positive components shall exist such that the components of $(I - H)u_0$ are all positive. The existence of such a vector u_0 follows automatically from known results on matrices if the eigenvalues of H all lie inside the unit circle and H is non-singular. We write $u_0 = Hu_0 + c + \epsilon$; then $c_j + \epsilon_j > 0$, $j = 1, \cdots, n$.

If the problem in view is to solve $Au = b$, then theoretically speaking, matrices H and M can always be selected so that the equivalent system $u = Hu + Mb$, with $H + MA = I$, satisfies the conditions imposed here. For example, choose H to be a principal diagonal matrix with elements lying between 0 and 1. Then its eigenvalues lie in the unit circle, and the existence of a vector u_0 with the required property is assured. (Of course, any particular a priori choice of H such as this usually means that M must be determined from $M = A^{-1}(I - H)$, which requires a knowledge of A^{-1}. In practice therefore it is M that will probably be chosen first, not H.)

We note that ϵ (or rather $-\epsilon$) is the "residual" in the classical theory of solutions of the system $(I - H)u = c$. It is easy to show that if the successive approximations $u_1, u_2, \cdots$ to the solution u of (10.2) are defined by the recursion relation (10.1), then

$$u_N - u = H^N(I - H)^{-1}\epsilon.$$

The second member of this equation is the truncation error (as opposed to the statistical error) in the statistical solution about to be proposed.

At this point we reformulate the problem in terms of the partitioned matrix and vectors introduced in Section 6. The equation (10.1) can be rewritten as

$$(u_{N+1} \vdots c) = H^* \times (u_N \vdots c) \tag{10.3}$$

and the equation (10.2) becomes

$$(u \vdots c) = H^* \times (u \vdots c). \tag{10.4}$$

The problem of solving (10.2) then becomes one of finding an eigenvector of H^* for the eigenvalue unity, given preassigned values $c_1, c_2, \cdots, c_n$ for the last n of the $2n$ components of the eigenvector. This formulation of the "steady state" problem as an eigenvalue problem permits us to make an approach to the zero-variance sampling situation discussed earlier in the section.

We define correspondences between the points $x_{n+1}, x_{n+2}, \cdots, x_{2n}$, and the components of ϵ and c respectively by $c(x_{i+n}) = c_i$, $\epsilon(x_{i+n}) = \epsilon_i$, $i = 1, \cdots, n$; and we introduce the vector c^* whose components are given by

$$c_i^* = c^*(x_i) = \begin{cases} u_0(x_i), & i = 1, 2, \cdots, n \\ c(x_i) + \epsilon(x_i), & i = n + 1, \cdots, 2n. \end{cases}$$

That is, $c^* = (u_0 \vdots c + \epsilon)$. The n equations represented by $u_0 = Hu_0 + c + \epsilon$ which define the components of ϵ are equivalent to the $2n$ equations represented by

$$c^* = H^*c^*. \tag{10.5}$$

We are now ready to set up the basic random walk, and the corresponding m-chain whose mean value is the solution of (10.1).

The transition probabilities of the random walk will be given by

$$p_{ij}^* = h_{ij}^* [c_j^*/c_i^*]$$

It is to be noted that $p_{ij}^* \geqq 0$, and because of (10.5),

$$\sum_{j=1}^{2n} p_{ij}^* = 1, \qquad i = 1, \cdots, 2n$$

Then z_{ij}^* is chosen so that

$$z_{ij}^* = c_i^*/c_j^*$$

Letting the random variables $X_0^*, X_1^*, \cdots$ denote as usual the successive states visited by a random walk having these transition probabilities, and substituting into $Z_N^*[(u_0 : c), z^*]$, we obtain

$$Z_N^* = \frac{c^*(X_0^*)}{c^*(X_1^*)} \cdot \frac{c^*(X_1^*)}{c^*(X_2^*)} \cdots \frac{c^*(X_{N-1}^*)}{c^*(X_N^*)} \times \begin{cases} u_0(X_N^*), \text{ if } X_N^* = x_i, i = 1, 2, \cdots, n, \\ c(X_N^*), \text{ if } X_N^* = x_i, i = n+1, \\ \qquad\qquad\qquad\qquad n+2, \cdots, 2n. \end{cases}$$

Cancelling terms and using the definition of c^*, we find that for X_0^* on $x_1, \cdots, x_n$,

$$Z_N^* = u_0(X_0^*)\left[1 - \delta(X_N^*)\frac{\epsilon(X_N^*)}{c(X_N^*) + \epsilon(X_N^*)}\right], \tag{10.6}$$

where

$$\delta(x_i) = \begin{cases} 0, & i = 1, \cdots, n \\ 1, & i = n+1, \cdots, 2n. \end{cases}$$

It is known from the results of Section 6 that $E(Z_N^* \mid X_0 = x_i) = u_{Ni}, i \leqq n$, where u_N is defined by (10.1) above.

The practical procedure implied here for estimating u_N is simply this: Start a random walk at x_i, using the transition probabilities p_{ij}^*. If absorption has not taken place after N steps (that is, if the random walk has not reached any of the points $x_{n+1}, x_{n+2}, \cdots, x_{2n}$), then record u_{0i}. If absorption *does* take place during the N steps, stop the walk then and there, note the index i of the last point x_i touched before absorption, and record $u_{0i}c_i/(c_i + \epsilon_i)$. Do this for many random walks, and then average up the recorded values.

We now consider the statistical error of the procedure. From (10.6),[19]

$$\operatorname{Var}(Z_N^* \mid X_0^* = x_i) = u_{0i}^2 \operatorname{Var}\left[\delta(X_N^*)\frac{\epsilon(X_N^*)}{c(X_N^*) + \epsilon(X_N^*)} \,\middle|\, X_0^* = x_i\right], \qquad i = 1, \cdots, n.$$

(We note in passing that if $\epsilon = 0$, then the variance vanishes, as it should.) Proceeding as we did in Section 4, and letting

$$e = \max_i \frac{\epsilon_i}{c_i + \epsilon_i},$$

[19] By the symbol $\operatorname{Var}(Y \mid b)$, we mean the variance of the conditional distribution of Y, given that the event b has occurred.

we get

$$\text{Var}\,(Z_N^* \mid X_0^* = x_i) < \tfrac{1}{2}\epsilon^2 u_{0i}^2 \qquad i = 1, \cdots, n,$$

which for practical purposes is about as satisfactory an appraisal of the variance as it seems possible to obtain. It contains none of the unknown quantities in the problem.

By using the formulas and methods of Section 7, it is easy to get an explicit formula for the variance. The result is

$$\text{Var}\,(Z_N^* \mid X_0^* = x_i)$$
$$= u_{0i}(I - H^N)(I - H)^{-1}\,\epsilon']_i - \{(I - H)^{-1}\,\epsilon]_i\}^2, \qquad i = 1, \cdots, n,$$

where

$$\epsilon_i' = \frac{\epsilon_i^2}{c_i + \epsilon_i}, \qquad i = 1, \cdots, n.$$

One somewhat interesting conclusion that can be drawn from this formula is that if u_0 is chosen so as to have constant components, then the variance will have one standard limiting vector as $N \to \infty$, no matter what the magnitude of the components of u_0 may be. It turns out that this limiting vector is

$$(I - H)^{-1}c' - u^2,$$

where u is the solution of (10.2) and

$$c_i' = \frac{c_i^2}{\sum_{j=1}^n (I - H)_{ij}}, \qquad i = 1, \cdots, n.$$

REFERENCES

1. Courant, R., Friedrichs, K., and Lewy, H., Uber die Partiellen Differenzengleichungen der Mathematischen Physik, Math. Annalen, vol. 100 (1948), pp. 32-74.
2. Courant, R., and Hilbert, D., Methoden der Mathematischen Physik, Berlin, 1931.
3. Cramer, H., Mathematical Methods of Statistics, Princeton, 1946.
4. Curtiss, J. H., Sampling Methods Applied to Differential and Difference Equations. Proceedings of a Seminar on Scientific Computation, held by the International Business Machines Corporation, Endicott, New York, November, 1949. Pp. 87–109.
5. Cutkosky, R. E., A Monte Carlo method for solving a class of integral equations, Journal of Research of the National Bureau of Standards.
6. Deming, W. Edwards, Some Theory of Sampling, New York, 1950.
7. Frazer, R. A., Duncan, W. J., and Collar, A. R., Elementary Matrices, Cambridge, 1950.
8. Feller, W., An Introduction to Probability Theory and its Applications, New York, 1950.
9. Forsythe, G. E., Tentative Classification of methods and bibliography on solving systems of linear equations. To appear in Simultaneous Equations and the Determination of Eigenvalues—Proceedings of a National Bureau of Standards Symposium held in Los Angeles, August, 1951.
10. Forsythe, G. E., and Leibler, R. A., Matrix Inversion by a Monte Carlo Method, Mathematical Tables and Other Aids to Computation, vol. 4 (1950), pp. 127–129.

11. KAHN, H., Stochastic (Monte Carlo) Attenuation Analysis, RAND Corp., Santa Monica, California. Report No. R-163, June 14, 1949.
12. KOLMOGOROV, A. N., Foundations of the Theory of Probability, New York, 1950.
13. MACDUFFEE, C. C., The Theory of Matrices, New York, 1946.
14. PROCEEDINGS OF A SYMPOSIUM ON THE MONTE CARLO METHOD, National Bureau of Standards Applied Mathematics Series No. 12, Washington, 1951.
15. MONTROLL, E. W., On the theory of Markov Chains, Annals of Math. Statistics, vol. 18 (1947), pp. 18–36.
16. WASOW, W., On the inversion of matrices by random walks, Mathematical Tables and Other Aids to Computation, vol. 6 (1951), pp. 78–81.
17. WASOW, W., On the mean duration of random walks, Journal of Research of the National Bureau of Standards, vol. 46 (1951), pp. 462–472.

NEW YORK UNIVERSITY

(Received March 23, 1953)

A GENERAL SOLUTION OF THE TWO-FREQUENCY MODULATION PRODUCT PROBLEM. I.[1]

By R. L. Sternberg and H. Kaufman

1. Introduction. The purpose of this paper is to present a general method for readily obtaining approximate numerical values of the amplitudes of the modulation products which occur in the output when a two-frequency input is applied to an arbitrary modulator having a continuous output versus input characteristic; a partial analytical solution of the problem is also indicated. In solving the problem it is shown, in particular, that approximate values of the modulation product amplitudes always can be determined as simple linear combinations of the values of four new functions introduced by W. R. Bennett [1, 2] and that the output can be approximated uniformly for all time by the convergent double Fourier series having these approximate values as coefficients. The approximate values of the modulation product amplitudes determined by the method given are shown to be capable of an arbitrary degree of refinement, subject only to limits of accuracy of determination of the four new functions mentioned, tables of which are to be given in Part II of the present paper, and an estimate of error is indicated. The method yields exact values of the modulation product amplitudes in the case of such modulators as the biased ideal rectifier or ideal limiter, whose output versus input characteristic is a continuous polygonal function, and is an extension of the method employed by Bennett [2] in connection with those two devices. Finally, the method will be applied in Part III to study in some detail a number of examples including among others the two particular modulators just mentioned.

The problem is formulated in §2. In §3 an approximate solution is derived. The coefficients in that solution are reduced in §4 by the method of Bennett and are thus found to be expressible in terms of the four new functions mentioned; several properties of these functions are noted and power series expansions for them are obtained in §5.

2. Formulation of the problem. Consider an arbitrary modulator whose output versus input characteristic $Y = Y(X)$ is a continuous function of X on a finite closed interval $-a \leqq X \leqq a$. Let the input $x(t)$ be given as

$$(2.1)\quad x(t) = P\cos(pt + \theta_p) + Q\cos(qt + \theta_q), \qquad 0 < P < P + Q \leqq 2P \leqq a.$$

Then without regard to convergence, the output $y(t) \equiv Y(x(t))$ has associated with it a double Fourier series of the form

$$(2.2)\quad y(t) \sim \tfrac{1}{2}\, C_{00} + {\sum_{m,n=0}^{*\infty}}\, C_{\pm mn}\cos(\omega_{\pm mn}t + \phi_{\pm mn}),$$

where the asterisk indicates that the summation is to be extended over both the upper and lower signs when $m \cdot n \neq 0$ and over only the upper sign when $m \cdot n = 0$

[1] Presented to the American Mathematical Society at their New York Meeting April 25, 1952.

with the single exception that the term occurring under the summation sign for $m = n = 0$ is not to be included in the sum. In (2.2) the angular frequencies and phase angles are given by the formulas

$$\omega_{\pm mn} = mp \pm nq, \qquad \phi_{\pm mn} = m\theta_p \pm n\theta_q, \qquad (m, n = 0, 1, 2, \cdots), \tag{2.3}$$

while in terms of the periodic continuous function of two variables

$$F(u, v) = Y(P \cos u + Q \cos v), \tag{2.4}$$

the coefficients in (2.2) are given by the formulas

$$C_{\pm mn} = (1/2\pi^2) \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} F(u, v) \cos (mu \pm nv)\, du\, dv, \quad (m, n = 0, 1, 2, \cdots). \tag{2.5}$$

The quantities $\frac{1}{2}C_{00}$ and $C_{\pm mn}$, $(m + n \geqq 1)$ are by definition the desired modulation product amplitudes; hence, our problem is to determine approximate numerical values of these quantities for given values of P, Q and modulator characteristics $Y = Y(X)$. By ordinary methods the problem is usually a laborious one since the integrals on the right in (2.5) are not usually expressible in terms of known functions and, moreover, have to be treated separately for each different modulator characteristic. In the balance of the paper we show how to solve this problem by an approximation process which yields values of the modulation product amplitudes $\frac{1}{2}C_{00}$ and $C_{\pm mn}$, $(m + n \geqq 1)$ accurate to within an arbitrary $4\epsilon/\pi$, expressed as simple linear combinations of integrals which are independent of the modulator characteristic and which for all values of m and n, in turn can be expressed as simple linear combinations of the values of the four new functions of Bennett [1, 2] mentioned in the introduction and to be tabulated in Part II. In this process it is shown, moreover, that the double Fourier series obtained from (2.2) by replacing the quantities $\frac{1}{2}C_{00}$ and $C_{\pm mn}$, $(m + n \geqq 1)$ by their approximate values converges and approximates the output $y(t)$ to within ϵ for all t independent of the convergence properties of the series (2.2) itself.

It may be remarked that although the approximation process may appear somewhat complicated, it will be found in practice to be quite simple to carry out.

3. Approximate solution. Let the modulator characteristic $Y = Y(X)$ be approximated to within an arbitrarily preassigned $\epsilon > 0$ uniformly on the finite closed interval $-a \leqq X \leqq a$ by a continuous polygonal function $S(X; \epsilon)$ expressed as a sum of the form

$$S(X; \epsilon) = Y(-a) + \sum_{j=1}^{N(\epsilon)} g_j Y(X; X_j) \tag{3.1}$$

where the X_j and g_j are suitable numbers with $-a \leqq X_1 < X_2 < \cdots < X_{N(\epsilon)} < a$ and the $Y(X; X_j)$ are unit ramp functions satisfying $Y(X; X_j) = 0$ for $X < X_j$ and $Y(X; X_j) = X - X_j$ for $X \geqq X_j$, $(j = 1, 2, \cdots, N(\epsilon))$. In view of the continuity and consequent uniform continuity of $Y(X)$ on $-a \leqq X \leqq a$ such an approximation can always be carried out and in practice a suitable choice

of the numbers X_j, g_j and $N(\epsilon)$ for a given ϵ can usually be made conveniently by graphical means. Thus in a particular case one may plot the modulator characteristic $Y = Y(X)$, draw in a polygonal curve approximating $Y = Y(X)$ to within ϵ uniformly on $-a \leqq X \leqq a$ and then determine the numbers X_j, g_j and $N(\epsilon)$ essentially by inspection from the polygonal curve. By Hobson [3; p. 710] or Tonelli [5; p. 454] each of the periodic continuous functions of two variables

$$f_j(u, v) = Y(P \cos u + Q \cos v; X_j), \qquad (j = 1, 2, \cdots, N(\epsilon)), \tag{3.2}$$

$$f(u, v; \epsilon) = Y(-a) + \sum_{j=1}^{N(\epsilon)} g_j Y(P \cos u + Q \cos v; X_j), \tag{3.3}$$

has a double Fourier series expansion everywhere convergent to that function. Moreover, introducing parameters $h_j = (X_j/P)$, $k = (Q/P)$ and factoring out the P in each function $f_j(u, v)$ these expansions can be written respectively in the forms

$$f_j(u, v) = \tfrac{1}{2}PA_{00}(h_j, k) + P\sum\nolimits_{m,n=0}^{*\infty} A_{\pm mn}(h_j, k) \cos(mu \pm nv), \qquad (j = 1, 2, \cdots, N(\epsilon)), \tag{3.4}$$

$$f(u, v; \epsilon) = \tfrac{1}{2}B_{00}(\epsilon) + \sum\nolimits_{m,n=0}^{*\infty} B_{\pm mn}(\epsilon) \cos(mu \pm nv), \tag{3.5}$$

where the coefficients are defined as

$$A_{\pm mn}(h_j, k) = (1/2\pi^2P) \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} f_j(u, v) \cos(mu \pm nv)\, du\, dv, \qquad (m, n = 0, 1, 2, \cdots; j = 1, 2, \cdots, N(\epsilon)) \tag{3.6}$$

$$B_{\pm mn}(\epsilon) = (1/2\pi^2) \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} f(u, v; \epsilon) \cos(mu \pm nv)\, du\, dv, \qquad (m, n = 0, 1, 2, \cdots)' \tag{3.7}$$

and may be seen with the aid of (3.2), (3.3) and an elementary theorem on integrals to be related by the formulas

$$\begin{aligned} \tfrac{1}{2}B_{00}(\epsilon) &= Y(-a) + \tfrac{1}{2}P\sum\nolimits_{j=1}^{N(\epsilon)} g_j A_{00}(h_j, k), \\ B_{\pm mn}(\epsilon) &= P\sum\nolimits_{j=1}^{N(\epsilon)} g_j A_{\pm mn}(h_j, k), \qquad (m + n \geqq 1). \end{aligned} \tag{3.8}$$

Formulas (3.8) constitute an approximate solution of our problem in terms of the coefficients $A_{\pm mn}(h_j, k)$ to be determined later; that is, the quantities $\tfrac{1}{2}B_{00}(\epsilon)$ and $B_{\pm mn}(\epsilon)$, $(m + n \geqq 1)$ as given by (3.8) are linear approximate expressions for the desired modulation product amplitudes $\tfrac{1}{2}C_{00}$ and $C_{\pm mn}$, $(m + n \geqq 1)$. Indeed, with the aid of the constructed properties of the function $f(u, v; \epsilon)$, particularly, the relation $|F(u, v) - f(u, v; \epsilon)| \leqq \epsilon$, to establish the inequalities

$$(1/2\pi^2) \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} |F(u, v) - f(u, v; \epsilon)| \cdot |\cos(mu \pm nv)|\, du\, dv \leqq \begin{cases} 2\epsilon, & (m = n = 0), \\ 4\epsilon/\pi, & (m + n \geqq 1), \end{cases} \tag{3.9}$$

it is seen readily from (2.5) and (3.7) that the approximate modulation product amplitudes given by (3.8) in terms of the coefficients (3.6) differ from the true values $\frac{1}{2}C_{00}$ and $C_{\pm mn}$, $(m + n \geqq 1)$ by not more than $4\epsilon/\pi$ in all cases, i.e., we have

$$|\tfrac{1}{2}B_{00}(\epsilon) - \tfrac{1}{2}C_{00}| \leqq \epsilon < 4\epsilon/\pi, |B_{\pm mn}(\epsilon) - C_{\pm mn}| \leqq 4\epsilon/\pi, \quad (m + n \geqq 1). \tag{3.10}$$

Moreover from the same properties of $f(u, v; \epsilon)$ and the convergence of its double Fourier series (3.5) it follows, putting $u = pt + \theta_p$, $v = qt + \theta_q$ in the latter, that the output $y(t) \equiv Y(x(t))$ is given to within ϵ for all t by the approximate formula

$$y(t) \cong \tfrac{1}{2}B_{00}(\epsilon) + \sum_{m,n=0}^{**\infty} B_{\pm mn}(\epsilon) \cos(\omega_{\pm mn}t + \phi_{\pm mn}). \tag{3.11}$$

Finally, in the case of a modulator whose characteristic $Y = Y(X)$ is a continuous polygonal function on the closed interval $-a \leqq X \leqq a$ it is clear that one may take $\epsilon = 0$ in the above process so that the modulation product amplitudes as given by (3.8), and the output as given by (3.11), are in this case exact.

It remains now to show how to determine the coefficients $A_{\pm mn}(h_j, k)$ defined by (3.6); we do this in §§4 and 5 below carrying out the reduction in §4 of the higher order coefficients $A_{\pm mn}(h_j, k)$ to linear combinations of the four lower order coefficients $A_{00}(h_j, k)$, $A_{10}(h_j, k)$, $A_{01}(h_j, k)$, $A_{11}(h_j, k)$ by the method of Bennett [2] and deriving in §5 power series expansions for the lower order coefficients mentioned.

4. Reduction of the coefficients. To simplify the coefficients $A_{\pm mn}(h_j, k)$ defined by the formulas (3.6) we begin, by noting from the symmetry properties of the functions $f_j(u, v)$, that $A_{+mn}(h_j, k) = A_{-mn}(h_j, k)$ for all m and n so that the double sign can hereafter be dropped. Next, noting the formulas

$$f_j(u, v) = \begin{cases} P(\cos u + k \cos v - h_j), \\ 0 \end{cases}, \quad \cos u + k \cos v \begin{cases} \geqq h_j, \\ < h_j, \end{cases} \quad (j = 1, 2, \cdots, N(\epsilon)), \tag{4.1}$$

and considering the various forms that can be assumed by the curve $\cos u + k \cos v = h$, $h = h_j$, which bounds the region in the closed square $-\pi \leqq u \leqq \pi$, $-\pi \leqq v \leqq \pi$ over which $f_j(u, v) \not\equiv 0$ and, hence, over which the integrations in (3.6) actually have to be carried out, we find that there are five cases corresponding to different values of the parameters $h = h_j = (X_j/P)$ and $k = (Q/P)$. These five cases are defined by the inequalities

$$\begin{gathered} \text{(o) } h - k \geqq 1, \quad \text{(i)} \begin{cases} h - k < 1, \\ h + k > 1, \end{cases} \quad \text{(ii)} \begin{cases} h - k \geqq -1, \\ h + k \leqq 1, \end{cases} \\ \text{(iii)} \begin{cases} h - k < -1, \\ h + k > -1, \end{cases} \quad (\infty)\ h + k \leqq -1, \end{gathered} \tag{4.2}$$

and have thus been chosen according as the curves of the family $\cos u + k \cos v = h$ in the closed square $-\pi \leqq u \leqq \pi$, $-\pi \leqq v \leqq \pi$, except for certain border-

line curves which separate cases, consist of: (o) at most the point $u = v = 0$ the corresponding function $f_j(u, v)$ being everywhere zero, (i) a single closed curve, (ii) two open branches, (iii) four open branches or (∞) at most the four points $u = \pm\pi$, $v = \pm\pi$, the corresponding function $f_j(u, v)$ being positive at least at all other points. For figures showing various forms of these curves we refer to Bennett [1; p. 231] and [2; p. 144]; in connection with the latter reference it may be noted that Bennett's case (I) corresponds to our case (iii) and vice versa.

We first dispose of the two degenerate cases (o) and (∞). In these cases (3.6) and (4.1) yield at once the results

$$\text{(o)}\quad A_{mn}(h, k) = 0, \qquad (m, n = 0, 1, 2, \cdots),$$

$$(\infty)\quad A_{00}(h, k) = -2h, \quad A_{10}(h, k) = 1, \quad A_{01}(h, k) = k, \quad A_{mn}(h, k) = 0, \qquad (m + n > 1), \tag{4.3}$$

where $h = h_j$. Proceeding now to the non-degenerate cases (i), (ii) and (iii) we find, with the aid of formulas (4.1), that in these cases the $A_{mn}(h, k)$ defined by (3.6) can always be expressed in either of the forms

$$\begin{aligned} A_{mn}(h, k) = {} & (2/\pi^2)\int_a^b \left[\int_0^{c(v)} (\cos u + k\cos v - h)\cos mu\, du\right]\cos nv\, dv \\ & + (2/\pi^2)\int_0^a \left[\int_0^{\pi} (\cos u + k\cos v - h)\cos mu\, du\right]\cos nv\, dv, \\ & \qquad (m, n = 0, 1, 2, \cdots), \end{aligned} \tag{4.4}$$

$$\begin{aligned} A_{mn}(h, k) = {} & (2/\pi^2)\int_a^\beta \left[\int_0^{\gamma(u)} (\cos u + k\cos v - h)\cos nv\, dv\right]\cos mu\, du \\ & + (2/\pi^2)\int_0^\alpha \left[\int_0^{\pi} (\cos u + k\cos v - h)\cos nv\, dv\right]\cos mu\, du, \\ & \qquad (m, n = 0, 1, 2, \cdots), \end{aligned} \tag{4.5}$$

where a, b, c and α, β, γ are defined in terms of the parameters $h = h_j = (X_j/P)$, $k = (Q/P)$, $h' = (h/k)$ and $k' = (1/k)$ in cases (i), (ii) and (iii) respectively as

$$\text{(i)}\ a = 0, \qquad b = \cos^{-1}(h' - k'), \quad \alpha = 0, \qquad \beta = \cos^{-1}(h - k),$$

$$\text{(ii)}\ a = 0, \qquad b = \pi, \qquad \alpha = \cos^{-1}(h + k), \qquad \beta = \cos^{-1}(h - k),$$

$$\text{(iii)}\ a = \cos^{-1}(h' + k'), \quad b = \pi, \qquad \alpha = \cos^{-1}(h + k), \qquad \beta = \pi,$$

and in all three cases we have $c(v) = \cos^{-1}(h - k\cos v)$ and $\gamma(u) = \cos^{-1}(h' - k'\cos u)$. In a number of subcases the second integrals in (4.4) and

(4.5) are zero; in particular, the second integral in (4.4) is zero in cases (i) and (ii) for all m, n and in case (iii) for $m \geqq 2$, while the second integral in (4.5) is zero in case (i) for all m, n and in cases (ii) and (iii) for $n \geqq 2$.

The inner integrations in each of the double integrals in (4.4) and (4.5) can always be performed. With the aid of this fact one can establish recursion relations between the inner integrals in the first double integrals of (4.4) and (4.5) separately, both resulting relations being required, which then together lead, after elementary but somewhat lengthy calculations, to the following set of linear recursion identities between the coefficients $A_{mn}(h, k)$ as functions of the parameters h and k, which are valid in all five cases of these parameters and are of particular interest in cases (i), (ii) and (iii):

$$
\begin{aligned}
(m - n + 3)A_{m+1,n-1} &\equiv -(m + n - 3)A_{m-1,n-1} \\
&\quad + 2mhA_{m,n-1} - 2mkA_{mn}, \qquad (m, n \geqq 1), \\
(m + n + 1)A_{mn} &\equiv -(m - n - 3)A_{m-2,n} - 2(m - 1)kA_{m-1,n-1} \\
&\quad + 2(m - 1)hA_{m-1,n}, \qquad (m \geqq 2, n \geqq 1), \\
(n + m + 1)A_{mn} &\equiv -(n - m - 3)A_{m,n-2} - 2(n - 1)k'A_{m-1,n-1} \\
&\quad + 2(n - 1)h'A_{m,n-1}, \qquad (m \geqq 1, n \geqq 2), \\
(n - m + 3)A_{m-1,n+1} &\equiv -(n + m - 3)A_{m-1,n-1} \\
&\quad + 2nh'A_{m-1,n} - 2nk'A_{mn}, \qquad (m, n \geqq 1),
\end{aligned}
\tag{4.6}
$$

where again $h' = (h/k)$ and $k' = (1/k)$. Relations equivalent to (4.6), except for two minor misprints, have been given by Bennett [2; p. 159] and are therein attributed to S. O. Rice. With the aid of relations (4.6) the higher order coefficients $A_{mn}(h, k)$ can be expressed in terms of the first four coefficients, i.e., in terms of $A_{00}(h, k)$, $A_{10}(h, k)$, $A_{01}(h, k)$ and $A_{11}(h, k)$. Thus for example we have the following formulas for the second and third order coefficients:

$$
\begin{aligned}
A_{20} &\equiv \tfrac{1}{3}A_{00} + \tfrac{2}{3}hA_{10} - \tfrac{2}{3}kA_{11}, \\
A_{02} &\equiv \tfrac{1}{3}A_{00} + \tfrac{2}{3}h'A_{01} - \tfrac{2}{3}k'A_{11}, \\
A_{30} &\equiv \tfrac{1}{3}hA_{00} + (\tfrac{1}{2}k^2 + \tfrac{2}{3}h^2)A_{10} - \tfrac{1}{2}kA_{01} - \tfrac{7}{6}hkA_{11}, \\
A_{21} &\equiv \tfrac{1}{2}A_{01} - \tfrac{1}{2}kA_{10} + \tfrac{1}{2}hA_{11}, \\
A_{12} &\equiv \tfrac{1}{2}A_{10} - \tfrac{1}{2}k'A_{01} + \tfrac{1}{2}h'A_{11}, \\
A_{03} &\equiv \tfrac{1}{3}h'A_{00} + (\tfrac{1}{2}k'^2 + \tfrac{2}{3}h'^2)A_{01} - \tfrac{1}{2}k'A_{10} - \tfrac{7}{6}h'k'A_{11}.
\end{aligned}
\tag{4.7}
$$

Assuming for the moment that values of the functions $A_{00}(h, k)$, $A_{10}(h, k)$, $A_{01}(h, k)$ and $A_{11}(h, k)$ for suitable values of the parameters h and k are known, one now may obtain readily an approximate numerical solution of the problem. To recapitulate briefly one proceeds as follows: (a) construct the polygonal approximating sum (3.1) graphically thereby determining the numbers X_j, g_j and $N(\epsilon)$, (b) form the sums (3.8) giving linear approximate expressions $\frac{1}{2}B_{00}(\epsilon)$

and $B_{\pm mn}(\epsilon)$, $(m + n \geqq 1)$ for the desired modulation product amplitudes $\frac{1}{2}C_{00}$ and $C_{\pm mn}$, $(m + n \geqq 1)$ and (c) determine the coefficients $A_{\pm mn}(h_j, k) = A_{mn}(h_j, k)$ in these approximate expressions where $h_j = (X_j/P)$ and $k = (Q/P)$, from formulas (4.3), (4.6) and (4.7) as linear combinations of the quantities $A_{00}(h_j, k)$, $A_{10}(h_j, k)$, $A_{01}(h_j, k)$ and $A_{11}(h_j, k)$. To realize an actual approximate numerical solution computation of the values of the functions $A_{00}(h, k)$, $A_{10}(h, k)$, $A_{01}(h, k)$ and $A_{11}(h, k)$ for a number of values of their parameters h and k has been undertaken and tables of these functions will be given in Part II of the present paper.

In certain cases, particularly in the case of modulators whose characteristic $Y = Y(X)$ satisfies the antisymmetric condition $Y(-X) = -Y(X)$, e.g., the symmetric case of the ideal limiter mentioned in the introduction and to be studied in detail in Part III, the above process can be somewhat simplified by a symmetric choice of the numbers X_j and corresponding antisymmetric choice of the numbers g_j in (3.1) together with use of the following identities which again are valid in all five cases of the parameters h and k:

$$A_{00}(-h, k) \equiv A_{00}(h, k) + 2h, \qquad A_{10}(-h, k) \equiv 1 - A_{10}(h, k),$$

$$A_{01}(-h, k) \equiv k - A_{01}(h, k), \qquad A_{mn}(-h, k) \equiv (-1)^{m+n} A_{mn}(h, k), \tag{4.8}$$

$$(m + n > 1).$$

These relations also have been given by Bennett [2; p. 158] and may be established in all five cases by elementary means, the essential feature of the derivation in cases (i), (ii) and (iii) being to note that if h is replaced by $-h$ in the equation $\cos u + k \cos v = h$ then a curve in case (i) is carried into a curve in case (iii) and vice versa, while a curve in case (ii) is carried into another curve in the same case provided in the latter instance that $h \neq 0$. Relations (4.8) have been used extensively in constructing the tables of the functions $A_{00}(h, k)$, $A_{10}(h, k)$, $A_{01}(h, k)$ and $A_{11}(h, k)$ for Part II and, if desired, may also be used to simplify slightly the derivation of the identities (4.6) but are not essential to that derivation.

5. Power series for $A_{00}(h, k)$, $A_{10}(h, k)$, $A_{01}(h, k)$ and $A_{11}(h, k)$. For values of h and k in case (ii), i.e., for $|h| + k \leqq 1$, one may readily obtain double series in powers of h and k for the functions in the title. To this end we begin by noting with the aid of (4.4) in case (ii) the formulas

$$\begin{aligned} A_{00}(h, k) &= (2/\pi^2) \int_0^\pi R(h - k \cos v)\, dv, \\ A_{10}(h, k) &= (1/\pi^2) \int_0^\pi S(h - k \cos v)\, dv, \\ A_{01}(h, k) &= (2/\pi^2) \int_0^\pi R(h - k \cos v) \cos v\, dv, \\ A_{11}(h, k) &= (1/\pi^2) \int_0^\pi S(h - k \cos v) \cos v\, dv, \end{aligned} \tag{5.1}$$

where

$$(5.2)\quad R(w) = (1 - w^2)^{1/2} - w\cos^{-1}w, \qquad S(w) = \cos^{-1}w - w(1 - w^2)^{1/2}.$$

Next noting that in case (ii) we have $|h - k\cos v| \leqq |h| + k \leqq 1$ we expand each of the functions $R(w)$ and $S(w)$ in powers of w and note with the aid of Knopp [4; p. 215 and p. 426] and elementary arguments that the resulting series are not only convergent but are uniformly convergent for $|w| \leqq 1$. Hence, putting $w = h - k\cos v$ and substituting the series for $R(h - k\cos v)$ and $S(h - k\cos v)$ in the integrands of the integrals in (5.1), we apply the binomial theorem to expand powers of $(h - k\cos v)$ and integrate termwise to obtain the expansions

$$\begin{aligned}(5.3)\quad \frac{\pi}{2}A_{00}(h,k) + \frac{\pi}{2}h &= 1\binom{0}{0} + \frac{1}{2\cdot 1}\binom{2}{0}h^2 + \frac{1}{2\cdot 4\cdot 3}\binom{4}{0}h^4 + \cdots \\ &+ \frac{1}{2\cdot 1}\binom{2}{2}\frac{1}{2}k^2 + \frac{1}{2\cdot 4\cdot 3}\binom{4}{2}\frac{1}{2}h^2k^2 + \cdots + \frac{1}{2\cdot 4\cdot 3}\binom{4}{4}\frac{1\cdot 3}{2\cdot 4}k^4 + \cdots\end{aligned}$$

$$\begin{aligned}(5.4)\quad \frac{\pi}{2}A_{10}(h,k) - \frac{\pi}{4} &= -1\binom{1}{0}h + \frac{1}{2\cdot 3}\binom{3}{0}h^3 + \frac{1}{2\cdot 4\cdot 5}\binom{5}{0}h^5 + \cdots \\ &+ \frac{1}{2\cdot 3}\binom{3}{2}\frac{1}{2}hk^2 + \frac{1}{2\cdot 4\cdot 5}\binom{5}{2}\frac{1}{2}h^3k^2 + \cdots + \frac{1}{2\cdot 4\cdot 5}\binom{5}{4}\frac{1\cdot 3}{2\cdot 4}hk^4 + \cdots\end{aligned}$$

$$\begin{aligned}(5.5)\quad \frac{\pi}{2}A_{01}(h,k) - \frac{\pi}{4}k &= -\frac{1}{2\cdot 1}\binom{2}{1}\frac{1}{2}hk - \frac{1}{2\cdot 4\cdot 3}\binom{4}{1}\frac{1}{2}h^3k \\ &- \frac{1\cdot 3}{2\cdot 4\cdot 6\cdot 5}\binom{6}{1}\frac{1}{2}h^5k - \cdots - \frac{1}{2\cdot 4\cdot 3}\binom{4}{3}\frac{1\cdot 3}{2\cdot 4}hk^3 \\ &- \frac{1\cdot 3}{2\cdot 4\cdot 6\cdot 5}\binom{6}{3}\frac{1\cdot 3}{2\cdot 4}h^3k^3 - \cdots - \frac{1\cdot 3}{2\cdot 4\cdot 6\cdot 5}\binom{6}{5}\frac{1\cdot 3\cdot 5}{2\cdot 4\cdot 6}hk^5 - \cdots\end{aligned}$$

$$\begin{aligned}(5.6)\quad \frac{\pi}{2}A_{11}(h,k) &= 1\binom{1}{1}\frac{1}{2}k - \frac{1}{2\cdot 3}\binom{3}{1}\frac{1}{2}h^2k - \frac{1}{2\cdot 4\cdot 5}\binom{5}{1}\frac{1}{2}h^4k - \cdots \\ &- \frac{1}{2\cdot 3}\binom{3}{3}\frac{1\cdot 3}{2\cdot 4}k^3 - \frac{1}{2\cdot 4\cdot 5}\binom{5}{3}\frac{1\cdot 3}{2\cdot 4}h^2k^3 - \cdots \\ &- \frac{1}{2\cdot 4\cdot 5}\binom{5}{5}\frac{1\cdot 3\cdot 5}{2\cdot 4\cdot 6}k^5 - \cdots\end{aligned}$$

each of which converges for $|h| + k \leqq 1$, i.e., for h and k in case (ii) and where the binomial symbols are defined as in Knopp [4; p. 209]. The coefficients in these series follow a pattern which may be easily determined by checking through the details of the derivations; in particular, in each coefficient the left hand factor arises from the expansion of the relevant function $R(w)$ or $S(w)$ in (5.2), the middle factor arises from the binomial development of the powers of $w =$

$h - k\cos v$ in the expansions just mentioned and the right hand factor results from carrying out the integration.

For $|h| + k < 1$ one may establish readily a convenient estimate for the error which results when the series (5.3), (5.4), (5.5) and (5.6) are truncated after a finite number of terms. Thus by employing a suitable geometric series to estimate the remainder after the term in w^{2n} in the expansion of $R(w)$ or the term in w^{2n+1} in the expansion of $S(w)$ for $|w| < 1$, one may show that if the series (5.3) and (5.5) are broken off immediately after the terms in h and k of total degree $2n$, or the series (5.4) and (5.6) are broken off immediately after the terms in h and k of total degree $2n + 1$, then for $|h| + k < 1$ the error $E(h, k)$ so incurred in the evaluation of the functions $A_{00}(h, k)$, $A_{10}(h, k)$, $A_{01}(h, k)$ and $A_{11}(h, k)$ by the resulting finite approximations will satisfy

$$(5.7)\quad |E(h,k)| \leqq \frac{1}{\pi^2}\left[\frac{3\cdot 5\cdots(2n-1)}{4\cdot 6\cdots(2n+2)(2n+1)}\right]\int_0^\pi \left[\frac{(h-k\cos v)^{2n+2}}{1-(h-k\cos v)^2}\right] dv.$$

For practical purposes the integral in (5.7) may either be estimated with the aid of the relation $|h - k\cos v| \leqq |h| + k < 1$ or may be evaluated numerically.

The series (5.3), (5.4), (5.5) and (5.6) provide a convenient method for evaluation of the functions $A_{00}(h, k)$, $A_{10}(h, k)$, $A_{01}(h, k)$ and $A_{11}(h, k)$ in case (ii) and have been employed for that purpose together with specially derived numerical integration formulas in the preparation of the tables for Part II; the series (5.3), (5.4), (5.5) and (5.6) may also be combined with the aid of the identities (4.6) and (4.7) to yield similar series for the higher order functions $A_{mn}(h, k)$. Moreover, to return to our modulation product problem, the series (5.3), (5.4), (5.5) and (5.6) together with relations (4.3), (4.6), (4.7) and (4.8) can be caused to yield a partial analytical solution for the desired modulation product amplitudes in the case of certain simple modulators having especially simple continuous polygonal characteristics such as the biased ideal rectifier or ideal limiter mentioned in the introduction and to be studied in detail in Part III. Thus for the symmetric ideal limiter whose characteristic $Y = Y(X)$ is polygonal and satisfies the antisymmetric condition $Y(-X) = -Y(X)$ we will obtain in Part III with the aid of the series and formulas mentioned, power series expansions for the quantities $\frac{1}{2}C_{00}$ and $C_{\pm mn}$, $(m + n \geqq 1)$ in powers of $h = (X_0/P)$ and $k = (Q/P)$ where $\pm X_0$ are the input values for which the limiter saturates and P and Q are as before the values of the input amplitudes in (2.1), the series being convergent provided only that $P - Q \geqq X_0 > 0$.

Acknowledgment. The authors wish to acknowledge their indebtedness to the Wright Patterson Air Development Center for partial contractual support of this work, to Dr. M. A. Meyer for suggesting a specific problem which led naturally to the general problem here considered, to Mrs. S. S. Fahlquist for extensive preliminary calculations performed in connection with the tables for Part II and to Miss Shirley Drell and Mrs. C. F. Ramsey for additional preliminary calculations.

REFERENCES

1. Bennett, W. R., "New Results in the Calculation of Modulation Products", *Bell System Technical Journal*, vol. 12 (1933), pp. 228–243.
2. Bennett, W. R., "The Biased Ideal Rectifier", *Bell System Technical Journal*, vol. 26 (1947), pp. 139–169.
3. Hobson, E. W., *The Theory of Functions of a Real Variable and Theory of Fourier's Series*, vol. 2, 2nd ed., Harren Press (Washington, D. C.) 1950.
4. Knopp, K., *Theory and Application of Infinite Series*, 2nd English ed., Blackie and Son (London and Glasgow) 1951.
5. Tonelli, L., *Serie Trigonometriche*, Nicola Zanichelli (Bologna) 1928.

Laboratory for Electronics, Inc.
Boston, Massachusetts.

(Received April 25, 1952)

ON RICHARDSON'S METHOD FOR SOLVING LINEAR SYSTEMS WITH POSITIVE DEFINITE MATRICES

By David Young[1]

1. Introduction. In 1910 L. F. Richardson [6][2] presented an iterative method for solving systems of linear equations which arise in the finite difference solutions of boundary value problems associated with elliptic partial differential equations. For a linear system of the form

$$\sum_{j=1}^{N} a_{i,j} u_j + d_i = 0, \qquad (i = 1, 2, \cdots, N) \tag{1}$$

where $u_1, u_2, \cdots, u_N$ are the unknowns, the iterative formula for Richardson's method is

$$u_i^{(n)} = u_i^{(n-1)} + \beta_n \left\{ \sum_{j=1}^{N} a_{i,j} u_j^{(n-1)} + d_i \right\}, \qquad (n \geq 1), \quad (i = 1, 2, \cdots, N) \tag{2}$$

where the trial values $u_1^{(0)}, u_2^{(0)}, \cdots, u_N^{(0)}$ are arbitrary and where the constants β_n are "relaxation factors" to be chosen.

It appears that even though Richardson's method was proposed more than forty years ago it had been used very little until quite recently. Richardson gave only rough suggestions for choosing the β_n, and if the choice of these numbers is based only on the suggestions given, then the rate of convergence is not much larger than that of the method wherein one uses a single fixed β_n for all iterations. Recently Flanders and Shortley, [2], used a method similar to Richardson's method for problems of finding by iteration the smallest eigenvalues of certain matrices. They used a theorem, proved by Markoff, [5], on Tschebyscheff polynomials to choose numbers analogous to the β_n of (2) and thereby obtained more rapid convergence.

This theorem on Tschebyscheff polynomials was applied to the solution of linear systems independently by Lanczos, [13], by Shortley, [14], and by the author, [12]. The method used by Lanczos differs from Richardson's method and does not appear to be quite as simple from a computational point of view, although the possible gain in rate of convergence should be the same. Shortley used Tschebyscheff polynomials in the solution of the difference equation analogue of Laplace's equation in two and three dimensions. He also obtained asymptotic estimates for the gain in the rate of convergence. His procedure, while ultimately equivalent to Richardson's method, differs in a practical computational aspect since it involves the use of linear operators which are polynomials in a single linear transformation defined on an N-dimensional vector space. In most cases the method requires the storage of many numbers and,

[1] This paper was prepared in part by research assigned to the Ballistic Research Laboratories, Aberdeen Proving Ground, Maryland, by the Office of the Chief of Ordnance under Project No. TB3-0007K. It was completed under contract DA-36-034-ORD-966, placed by the Office of Ordnance Research with the University of Maryland.

[2] Numbers in brackets, [ ], refer to the bibliography at the end of the paper.

although useful for computing machines with large storage capacities, it does not appear to be as promising as Richardson's method for use on faster machines having fairly limited storage capacities.

In the present paper we shall assume that the $N \times N$ matrix A of coefficients $a_{i,j}$ in (1) is symmetric and positive definite. We shall use Tschebyscheff polynomials to select the β_n which, in a certain sense, yield the fastest possible convergence for Richardson's method. If the β_n are suitably chosen, then the rate of convergence is asymptotically proportional to the square root of the rate of convergence of the method using the best single β_n, as the latter rate of convergence tends to zero. For a linear system associated with the usual difference analogue of the Dirichlet problem, the number of iterations varies inversely as the first power of the mesh size as compared to the second power of the mesh size for the method using a fixed β_n. Hence the gain in using Richardson's method is large for problems involving a fine mesh.

The best choice of the β_n depends on having good estimates of the maximum and minimum eigenvalues of A. Upper bounds present little difficulty, but lower bounds are often difficult to estimate. In Sec. 3 the dependence of the rate of convergence on the accuracy of the estimated lower bound of the eigenvalues of A is considered. The relative decrease in the rate of convergence is shown to be nearly proportional to the relative error in the estimated lower bound. In Sec. 4 the problem of the control of roundoff errors is discussed.

It is believed that the theorem on the gain in the rate of convergence for Richardson's method is new, as are the theorems on the effects of errors in estimating the minimum eigenvalue of A.

In Sec. 5 we give a brief comparison of Richardson's method and other methods including the Successive Overrelaxation Method and gradient methods. The order-of-magnitude gain in the convergence rate is the same for Richardson's method as for the Successive Overrelaxation Method, and Richardson's method can be shown to apply under more general conditions. On the other hand, the other method is better adapted for machines since it requires less storage, is simpler, has a much smaller tendency to accumulate roundoff errors, and in many cases it can be shown to converge at least twice as fast.

2. Convergence. Let $v = (v_1, v_2, \cdots, v_N)$ denote generically a vector in the vector space V of N - tuples of real numbers. Since the matrix A is symmetric and positive definite, there exist N positive eigenvalues $\nu_1, \nu_2, \cdots, \nu_N$ and N linearly independent eigenvectors $v^{(1)}, v^{(2)}, \cdots, v^{(N)}$ such that

$$\sum_{j=1}^{N} a_{i,j} v_j^{(k)} = \nu_k v_i^{(k)}, \qquad (i = 1, 2, \cdots, N), (k = 1, 2, \cdots, N). \tag{3}$$

Moreover, the eigenvectors are orthogonal under the inner product defined by

$$(x, y) = \sum_{i=1}^{N} x_i y_i$$

and hence form an orthogonal basis for V_N.

Since the determinant of A does not vanish there exists a unique solution of (1) which we denote by u. Let us define the error of the n-th iterated vector by

$$e^{(n)} = u^{(n)} - u. \tag{4}$$

Evidently, by (1), (2), and (4) we have

$$e_i^{(n)} = e_i^{(n-1)} + \beta_n \left\{ \sum_{j=1}^{N} a_{i,j} e_j^{(n-1)} \right\}, \qquad (i = 1, 2, \cdots, N) \tag{5}$$

or

$$e^{(n)} = Y(\beta_n)[e^{(n-1)}] \tag{5'}$$

where $Y(\beta_n)$ is a linear operator in V_N.

Given any $e^{(0)}$, there exist constants $c_1, c_2, \cdots, c_N$ such that

$$e^{(0)} = \sum_{k=1}^{N} c_k v^{(k)}. \tag{6}$$

Using (5) and (6) we obtain, for any integer p

$$e^{(p)} = \sum_{k=1}^{N} c_k v^{(k)} \prod_{n=1}^{p} (1 + \beta_n \nu_k), \tag{7}$$

and by the orthogonality of the eigenvectors we have

$$\| e^{(p)} \|^2 = (e^{(p)}, e^{(p)}) = \sum_{k=1}^{N} c_k^2 \| v^{(k)} \|^2 \left\{ \prod_{n=1}^{p} (1 + \beta_n \nu_k) \right\}^2, \tag{8}$$

or

$$\| e^{(p)} \|^2 \leqq \| e^{(0)} \|^2 \lambda_p^2 \tag{9}$$

where

$$\lambda_p = \operatorname{Max} \left| \prod_{n=1}^{p} (1 + \beta_n \nu_k) \right|, \qquad (k = 1, 2, \cdots, N). \tag{10}$$

For a given integer m, in order to make the least upper bound for all $e^{(0)} \in V_N$ of the ratio $\| e^{(m)} \| / \| e^{(0)} \|$ as small as possible we would choose the $\beta_n^{(m)}$ such that λ_m is a minimum. (Here the superscript m is introduced to indicate the dependence on m). However, if we know only that for some numbers a, b

$$0 < a \leqq \nu_k \leqq b, \qquad (k = 1, 2, \cdots, N) \tag{11}$$

then it appears natural to consider as the *optimum* set of $\beta_n^{(m)}$ those which minimize λ_m', where

$$\lambda_m' = \operatorname*{Max}_{a \leqq \nu \leqq b} \left| \prod_{n=1}^{m} (1 + \beta_n^{(m)} \nu) \right|. \tag{12}$$

Brauer [1] has shown that for all k

$$\nu_k \leqq \text{Max}\left\{\sum_{j=1}^{N} |a_{i,j}|\right\}, \qquad (i = 1, 2, \cdots, N);$$

hence we have an upper bound for the eigenvalues of A. There do not appear to be in the literature any explicit formulas for lower bounds, however.

For convenience, let us introduce the new variable γ defined by

$$\gamma = -\frac{2\nu}{b-a} + \frac{b+a}{b-a}. \tag{13}$$

The interval $a \leqq \nu \leqq b$ is mapped onto the interval $-1 \leqq \gamma \leqq 1$ in such a way that $\gamma = 1$ corresponds to $\nu = a$ and $\gamma = -1$ corresponds to $\nu = b$. Next, let $P_m(\nu)$ denote the polynomial

$$P_m(\nu) = \prod_{n=1}^{m} (1 + \beta_n^{(m)}\nu) \tag{14}$$

and let $Q_m(\gamma)$ be defined by

$$Q_m(\gamma) = P_m(\nu). \tag{15}$$

Since $P_m(0) = 1$ we have $Q_m[(b+a)/(b-a)] = 1$. The problem of minimizing λ_m' is thus equivalent to the problem of finding a polynomial in γ of degree m which equals unity for $\gamma = (b+a)/(b-a)$ and which has the smallest maximum absolute value in the interval $-1 \leqq \gamma \leqq 1$. By a theorem of Markoff [5], (A proof is given by Flanders and Shortley, [2].), the desired polynomial is $S_m(\gamma)$, where

$$S_m(\gamma) = T_m(\gamma)/T_m\left(\frac{b+a}{b-a}\right). \tag{16}$$

Here $T_m(\gamma)$ is the m-th order Tschebyscheff polynomial given by

$$T_m(\gamma) = \cos(m \cos^{-1}\gamma). \tag{17}$$

We note that (17) is valid even when γ is greater than unity, for although $\cos^{-1}\gamma$ is complex, $T_m(\gamma)$ is real. In fact, for $\gamma > 1$ we can replace (17) by

$$T_m(\gamma) = \tfrac{1}{2}[(\gamma + (\gamma^2 - 1)^{\frac{1}{2}})^m + (\gamma + (\gamma^2 - 1)^{\frac{1}{2}})^{-m}] = \cosh(m \cosh^{-1}\gamma). \tag{18}$$

In order to make $Q_m(\gamma) = S_m(\gamma)$, we equate corresponding zeros. Let

$$t_1^{(m)}, t_2^{(m)}, \cdots, t_m^{(m)}$$

denote the zeros of $T_m(\gamma)$; these are given by

$$t_n^{(m)} = \cos[(2n-1)\pi/2m], \qquad (n = 1, 2, \cdots, m). \tag{19}$$

Evidently the $t_n^{(m)}$ are the zeros of $S_m(\gamma)$. Next, the zeros of $Q_m(\gamma)$ are the values of γ corresponding to those values of ν which are the roots of the equation $P_m(\nu) = 0$. By (14) the roots of the latter equation are

$$\nu_n = -(\beta_n^{(m)})^{-1} \qquad (n = 1, 2, \cdots, m),$$

and by (13) the corresponding zeros of $S_m(\gamma)$ are given by

$$\gamma_n = \frac{2}{b-a}(\beta_n^{(m)})^{-1} + \frac{b+a}{b-a}. \tag{20}$$

Equating the zeros of $Q_m(\gamma)$ and $S_m(\gamma)$ and solving for $\beta_n^{(m)}$ we get

$$\beta_n^{(m)} = 2[(b-a)t_n^{(m)} - (b+a)]^{-1}, \quad (n = 1, 2, \cdots, m). \tag{21}$$

In the interval $-1 \leqq \gamma \leqq 1$, the maximum absolute value of $T_m(\gamma)$ equals unity; hence by (12), (14), (15), (16), and (18) we get

$$\lambda_m' = \max_{-1 \leqq \gamma \leqq 1} |S_m(\gamma)| = \left[T_m\left(\frac{b+a}{b-a}\right)\right]^{-1} < 1. \tag{22}$$

In the sense that λ_m' is minimized, the choice of the $\beta_n^{(m)}$ given by (21) is the best possible if exactly m iterations are to be performed. On the other hand, if after m iterations have been completed it is felt that more accuracy is needed, one can perform m' more iterations, where m' may equal m, using the appropriate $\beta_n^{(m')}$. However, although it would presumably have been more efficient to have used $\beta_n^{(m+m')}$ from the start, once the $\beta_n^{(m)}$ have been used, it is not efficient to do this since for $m'' > m$, very few of the $\beta_n^{(m)}$ will be included among the $\beta_n^{(m'')}$. It therefore appears best to choose a reasonable value of m and to use the $\beta_n^{(m)}$ in a cyclic order such as

$$\beta_n^{(m)} = \beta_r^{(m)} \tag{23}$$

where r is an integer such that $r \equiv n \pmod m$ and $1 \leqq r \leqq m$.

Since $\lambda_m \leqq \lambda_m'$ it follows from (9), (18) and (22) that with the choice of $\beta_n^{(m)}$ determined by (23), the method converges for any m. Incidentally, if $0 > \beta_n > -2/b$ for all n, then the method converges, since for $0 < \nu \leqq b$, we have

$$|1 + \beta_n \nu| < 1.$$

However, this condition is by no means necessary for convergence.

We remark that, although once a value of m has been chosen, the $\beta_n^{(m)}$ are determined by (21), nevertheless the order in which the $\beta_n^{(m)}$ are used within a cycle of m iterations is arbitrary. This ordering does not affect the *theoretical* rate of convergence, but, as we shall see in Sec. 4, it may be important because of the growth of roundoff errors.

3. Rate of convergence. We define the *rate of convergence* of a linear transformation T on the vector space V_N by

$$R(T) = -\log \Lambda \tag{24}$$

where Λ is the *spectral norm* of T, that is, Λ is the maximum of the absolute values of the eigenvalues of T. It is easy to show that if $\Lambda < 1$, then the rate of convergence of T is approximately inversely proportional to q, the number of times T must be applied to an arbitrary vector v so that $\| T^q[v] \|$ will be less

than a specified fraction of $\| v \|$, see for instance [11]. Thus if v were the initial error vector and an iteration scheme with transformation T were used repeatedly, the number of iterations needed to achieve a desired accuracy would be approximately inversely proportional to $R(T)$.

However, in Richardson's method, using the $\beta_n^{(m)}$ in a cyclic order, one does not use the same transformation repeatedly but, rather, a product of transformations. Therefore, it seems reasonable to consider the *average* rate of convergence

$$R_m = -(1/m) \log \Gamma_m \tag{25}$$

where Γ_m is the spectral norm of the transformation $\prod_{n=1}^{m} Y(\beta_n^{(m)})$. The $Y(\beta_n^{(m)})$ are defined by (5) and (5′).

Evidently, if v is an eigenvector of A with eigenvalue ν, then v is also an eigenvector of $Y(\beta)$ with eigenvalue $(1 + \beta\nu)$. Therefore, by (12) and (22) we have

$$\Gamma_m \leqq \left[T_m\left(\frac{b+a}{b-a}\right)\right]^{-1}, \tag{26}$$

and by (18) and (25)

$$R_m \geqq \frac{1}{m} \log T_m\left(\frac{b+a}{b-a}\right) = -\frac{1}{m} \log \frac{2\alpha^m}{1+\alpha^{2m}} \tag{27}$$

where, for convenience, we let

$$\alpha = \frac{\sigma}{1 + (1-\sigma^2)^{\frac{1}{2}}} \tag{28}$$

$$\sigma = \frac{b-a}{b+a}. \tag{29}$$

But since a is an eigenvalue of A, it follows from (13)–(17) that

$$\Gamma_m \geqq T_m[(b+a)/(b-a)]^{-1};$$

hence

$$R_m = -\frac{1}{m} \log \frac{2\alpha^m}{1+\alpha^{2m}} = \frac{1}{m} \log T_m(\sigma^{-1}). \tag{30}$$

For the method using a single relaxation factor, $m = 1$ and by (19) and (21) we have $\beta_n^{(1)} = -2/(b+a)$ for all n. The spectral norm of $Y(\beta_n^{(1)})$ is given by

$$\Gamma_1 = \sigma \tag{31}$$

and the rate of convergence is

$$R_1 = -\log \sigma. \tag{32}$$

Let us now compare R_m with R_1 as σ approaches unity.

THEOREM 1. Let r be an integer not less than 2 and let $m = m(\sigma)$ denote an integer such that $m \geqq \log 2^r/(-\log \alpha)$, where α is given by (28). If R_m and R_1 satisfy (30) and (32) respectively, then

$$\underline{\text{Lim}}_{\sigma \to 1-} (R_m/R_1^{\frac{1}{2}}) \geqq (1 - r^{-1})2^{\frac{1}{2}}. \tag{33}$$

PROOF. Since $m \geqq \log 2^r/(-\log \alpha)$ and $r \geqq 2$, it follows from (30) that $R_m \geqq (1 - r^{-1})(-\log \alpha)$. Therefore by (32) we have

$$\underline{\text{Lim}}_{\sigma \to 1-} (R_m/R_1^{\frac{1}{2}}) \geqq (1 - r^{-1}) \text{ Lim}_{\sigma \to 1-} [(-\log \alpha)/(-\log \sigma)^{\frac{1}{2}}].$$

But, using L'Hospital's rule and the identity

$$d\alpha/d\sigma = (1 - \sigma^2)^{-\frac{1}{2}}[1 + (1 - \sigma^2)^{\frac{1}{2}}]^{-1}$$

the limit in the right member of the last expression equals

$$2\text{Lim}_{\sigma \to 1-} [(-\log \sigma)^{\frac{1}{2}}/(1 - \sigma^2)^{\frac{1}{2}}] = 2\{\text{Lim}_{\sigma \to 1-} [(-\log \sigma)/(1 - \sigma^2)]\}^{\frac{1}{2}}$$

$$= 2\{\text{Lim}_{\sigma \to 1-} (2\sigma^2)^{-1}\}^{\frac{1}{2}} = 2^{\frac{1}{2}}$$

and the theorem follows.

For sufficiently large r, the right member of (33) can be made arbitrarily close to $2^{\frac{1}{2}}$. On the other hand, since at least m iterations must be performed, it would not be efficient to require that m be much larger than $-\log \rho/R_m$ where ρ is that fraction of $\| e^{(0)} \|$ to which the norm of the error must be reduced. This is true since $(-\log \rho)/R_m$ is the approximate number of iterations which would be required for convergence if each iteration affected the convergence equally. If $\rho \leqq 2^{-q}$, for some integer q, then r may be as large as $q + 1$ without m being excessive.

As an illustration, let us consider the set of linear equations one obtains in the numerical solution of the Dirichlet problem for the unit square by finite differences, see for instance [7], [8] or [11]. Let us assume that there are N interior net points and assign to each an integer i such that $1 \leqq i \leqq N$, and denote the value of the difference equation solution at the i-th point by u_i. Then in (1) we have $a_{i,i} = 4$ for all i and if $i \neq j$ then $a_{i,j} = 0$ or -1 depending on whether the points i, j are adjacent or non-adjacent. The d_i of (1) are functions of the boundary values, and do not affect the convergence of the iteration process.

It is not difficult to show, see for instance [11], that if h is the mesh size, then exact upper and lower bounds for the eigenvalues of A are, respectively, $a = 4(1 - \cos \pi h)$ and $b = 4(1 + \cos \pi h)$. Therefore by (29) we have $\sigma = \cos \pi h$ and $R_1 = -\log \sigma = \frac{1}{2}\pi^2 h^2 + O(h^4)$. If $m \geqq \log 2^r/(-\log \alpha)$, where

$$\alpha = \cos \pi h/(1 + \sin \pi h)$$

and $r \geqq 2$, then by Theorem 1 we have, as $h \to 0$

$$R_m \sim \tfrac{1}{2}\pi h.$$

Since the required number of iterations is approximately inversely proportional to the rate of convergence, it follows that the number of iterations varies as h^{-1} with Richardson's method using $m \geqq \log 2r/(-\log \alpha)$, (with $r \geqq 2$), as $h \to 0$, compared with h^{-2} using $m = 1$.

These results were found to hold reasonably well when a problem involving $h = 1/20$ was solved on the ORDVAC computing machine, (See [15]). The number of iterations required to satisfy a given convergence criterion with $m = 20$ was about one-ninth the number of iterations required with $m = 1$.

In general, however, we are not so fortunate as to have exact upper and lower bounds for the eigenvalues of A. As we have seen, lower bounds are the most difficult to obtain and we therefore proceed to discuss the effect on the rate of convergence of using a_1 in (21) to determine the $\beta_n^{(m)}$ instead of the true value a.

Let C_σ denote the set of symmetric positive definite matrices such that if $A \in C_\sigma$ then $\sigma = (b - a)/(b + a)$, where a, b denote respectively the minimum and maximum eigenvalues of A. Further, for any matrix $A \in C_\sigma$ let $R_m(\sigma, \theta, A)$ denote the rate of convergence of Richardson's method for a linear system with matrix A such that the m relaxation factors $\beta_n^{(m)}$ are determined from (21) using the correct value of b but replacing a by θa. Evidently by (30) we have

$$R_m(\sigma, 1, A) = (1/m) \log T_m(\sigma^{-1}). \tag{34}$$

First let us consider the case $\theta \leqq 1$. If $a_1 = \theta a$ then all eigenvalues of A lie in the range $a_1 \leqq \nu \leqq b$; hence we have

$$R_m(\sigma, \theta, A) \geqq (1/m) \log T_m(\sigma_1^{-1}), \qquad (\theta \leqq 1), \tag{35}$$

where $\sigma_1 = (b - a_1)/(b + a_1)$. We now prove

THEOREM 2. If $m = m(\sigma)$ is an integer such that $m \geqq \log 4/(-\log \alpha)$, where α is given by (28), and if $\theta \leqq 1$, then

$$\overline{\lim_{\sigma \to 1-}} \left(1 - \frac{R_m(\sigma, \theta, A)}{R_m(\sigma, 1, A)}\right) \leqq (1 - \theta)\theta^{\frac{1}{2}}. \tag{36}$$

PROOF. From (34) and (35) it follows that $R_m(\sigma, 1, A) - R_m(\sigma, \theta, A) \leqq (1/m)[\log T_m(\sigma^{-1}) - \log T_m(\sigma_1^{-1})]$. Let $c_1 = (1 - \sigma_1)/(1 + \sigma_1)$ and $c = (1 - \sigma)/(1 + \sigma)$. By the mean value theorem we have

$$\frac{1}{m} \log T_m(\sigma^{-1}) - \frac{1}{m} \log T_m(\sigma_1^{-1}) = (1 - \theta) \left\{\frac{d}{d\theta}\left[\frac{1}{m} \log T_m(\sigma_1^{-1})\right]\right\}_{\theta = \bar{\theta}}$$

where $\theta < \bar{\theta} < 1$. But

$$-\frac{d}{d\theta}\left[\frac{1}{m} \log T_m(\sigma_1^{-1})\right] = \frac{2c}{(1 - c_1)^2} \tanh [m \cosh^{-1}(\sigma_1^{-1})][\sigma_1^{-2} - 1]^{-\frac{1}{2}}$$

$$\leqq \frac{c^{\frac{1}{2}}}{1 - c_1}\left(\frac{c}{c_1}\right)^{\frac{1}{2}} = \frac{c^{\frac{1}{2}}}{1 - c_1}\theta^{-\frac{1}{2}}.$$

Therefore

$$\frac{1}{m}\log T_m(\sigma^{-1}) - \frac{1}{m}\log T_m(\sigma_1^{-1}) \leqq (1-\theta)\frac{c^{\frac{1}{2}}}{1-\bar{c}_1}\theta^{-\frac{1}{2}} \tag{37}$$

where $c_1 \leqq \bar{c}_1 \leqq c$. Moreover, since $m \geqq \log 4/(-\log \alpha)$, it follows that

$$R_m(\sigma, 1, A) \geqq -\tfrac{1}{2}\log \alpha.$$

Also, by using L'Hospital's rule we get $\operatorname{Lim}_{\sigma\to 1-}[(-\log\alpha)/2c^{\frac{1}{2}}] = 1$; hence if we divide both sides of (37) by $m^{-1}\log T_m(\sigma^{-1})$ and take the limit superior, we obtain the desired result.

Next, let us consider the case where $\theta \geqq 1$. The rate of convergence is equal to $-\log \Gamma_m$, where $\Gamma_m = \operatorname{Max}\,|\prod_{n=1}^{m}(1+\bar{\beta}_n^{(m)}\nu)|$ the maximum being taken over all eigenvalues ν of A, and where $\bar{\beta}_n^{(m)} = 2[(b-a_1)t_n^{(m)} - (b+a_1)]^{-1}$ with $a_1 = \theta a$. Letting $\gamma_1 = (-2\nu + b + a_1)(b-a_1)^{-1}$ we obtain as before

$$\prod_{n=1}^{m}(1+\bar{\beta}_n^{(m)}\nu) = T_m(\gamma_1)/T_m(\sigma_1^{-1}).$$

But since $\gamma_1(a) = 1 + 2(a_1 - a)/(b - a_1)$, and since $T_m(x)$ is an increasing function for $x \geqq 1$, it follows that $\Gamma_m = T_m(\gamma_1(a))/T_m(\sigma_1^{-1})$ and that

$$R_m(\sigma, \theta, A) = \frac{1}{m}\left[\log T_m(\sigma_1^{-1}) - \log T_m\left(1 + \frac{2(c_1 - c)}{1 - c_1}\right)\right], \qquad (\theta \geqq 1), \tag{38}$$

where $c = a/b = (1-\sigma)/(1+\sigma)$ and $c_1 = a_1/b = (1-\sigma_1)/(1+\sigma_1)$. We now prove

THEOREM 3. If $m = m(\sigma)$ is an integer such that $\log 4 \leqq m/(-\log\alpha) \leqq K$, where K is a constant, and if $\theta \geqq 1$, then

$$\overline{\operatorname{Lim}}_{\sigma\to 1-}\left(1 - \frac{R_m(\sigma, \theta, A)}{R_m(\sigma, 1, A)}\right) \leqq (\theta - 1)\left(K - \frac{15}{17}\theta^{-\frac{1}{2}}\right). \tag{39}$$

PROOF. For fixed m, we have by (18) and (38)

$$-\frac{d}{d\theta}[R_m(\sigma, \theta, A)] = \tanh\left[m\cosh^{-1}\left(1 + \frac{2(c_1 - c)}{1 - c_1}\right)\right]\frac{2c(1-c)}{(1-c_1)^2}$$

$$\cdot\left[\left(1 + \frac{2(c_1 - c)}{1 - c_1}\right)^2 - 1\right]^{-\frac{1}{2}} - \tanh\,[m\cosh^{-1}(\sigma_1^{-1})]\frac{2c\sigma_1^{-1}}{(1+c_1)^2}(1-\sigma_1^2)^{-\frac{1}{2}}.$$

Since $\tanh x \leqq x$ and $\log(1+x) \leqq x$ for $x \geqq 0$, we have

$$-\frac{d}{d\theta}[R_m(\sigma, \theta, A)] \leqq m\left[1 + \left(\frac{c_1 - c}{1 - c_1}\right)^{\frac{1}{2}}\right]\frac{2c(1-c)}{(1-c_1)^2}$$

$$- \tanh\,[m\cosh^{-1}(\sigma_1^{-1})]\frac{c}{(1+c_1)c_1^{\frac{1}{2}}\sigma_1}.$$

But $\overline{\operatorname{Lim}}_{\sigma\to 1-} mc^{\frac{1}{2}} \leqq \operatorname{Lim}_{\sigma\to 1-} [Kc^{\frac{1}{2}}/(-\log \alpha)] = K/2$, as in the proof of Theorem 2. Therefore, using the mean value theorem we get

$$\overline{\operatorname{Lim}}_{\sigma\to 1-}\left(1 - \frac{R_m(\sigma, \theta, A)}{R_m(\sigma, 1, A)}\right) \leqq \overline{\operatorname{Lim}}_{\sigma\to 1-}\left[(\theta - 1)\frac{c^{\frac{1}{2}}}{R_m(\sigma, 1, A)}\left\{\frac{2mc^{\frac{1}{2}}(1-c)}{(1-\bar{c}_1)^2}\right.\right.$$

$$\left.\left.\cdot\left[1 + \left(\frac{\bar{c}_1 - c}{1 - \bar{c}_1}\right)^{\frac{1}{2}}\right] - \tanh\left[m\cosh^{-1}(\bar{\sigma}_1^{-1})\right](c/\bar{c}_1)^{\frac{1}{2}}(1+\bar{c}_1)^{-1}\bar{\sigma}_1^{-1}\right\}\right]$$

where $c \leqq \bar{c}_1 \leqq c_1$. Since $\cosh^{-1}(\bar{\sigma}_1^{-1}) \geqq \cosh^{-1}(\sigma^{-1}) = -\log \alpha$, we have $\tanh[m\cosh^{-1}(\bar{\sigma}_1^{-1})] \geqq \tanh(\log 4) = 15/17$. Passing to the limit and recalling that $\underline{\operatorname{Lim}}_{\sigma\to 1-}[R_m(\sigma, 1, A)/c^{\frac{1}{2}}] \geqq 1$, and $\underline{\operatorname{Lim}}_{\sigma\to 1-}(c/\bar{c}_1)^{\frac{1}{2}} \geqq \theta^{\frac{1}{2}}$, the theorem follows.

From Theorems 2 and 3 we may conclude that the relative decrease in the rate of convergence, due to using a_1 instead of a, is approximately proportional to the relative error in estimating a, and that the constant of proportionality is not large. This is true whether the error is the result of overestimation or of underestimation.

4. Effect of roundoff. On the majority of large automatic computing machines all quantities must have modulus less than unity. Unless a floating decimal, or binary, point routine is used, at the expense of a large part of the storage capacity of the machine, numbers with modulus greater than unity must be premultiplied by a suitable scaling factor. However if in a certain storage register, the sizes of the numbers vary over wide ranges, in order to insure that the largest number has modulus less than unity, few significant digits can be available for the smaller numbers. For this reason in our problem it is desirable that the $u_i^{(m)}$ should not vary too much in magnitude.

This situation is not easy to achieve, however, if the quantities $|\beta_n^{(m)}|$ are large, as in the case where a, the lower bound of the eigenvalues of A, is small. In fact, by (21) we find that $\beta_1^{(m)} \to -a^{-1}$ as $m \to \infty$. Thus in the example of the preceding section, if $h^{-1} = 20$, then $a = .0492$, and if $m = 20$ then $\beta_1^{(m)} = -16.28$. For larger m, $|\beta_1^{(m)}|$ increases but never exceeds 20.32.

Let us now consider the error $e^{(p)}$ given by (7). Although the product

$$\prod_{n=1}^{p}(1 + \beta_n^{(m)}\nu_k)c_k$$

which is theoretically equal to the coefficient of $v^{(k)}$ in the expansion

$$e^{(p)} = \textstyle\sum_{k=1}^{N} c_k^{(p)} v^{(k)}$$

tends to zero as $p \to \infty$, the actual coefficient $c_k^{(p)}$ may not tend to zero. For, because of roundoff there exists some positive δ such that, in general, the actual coefficient will exceed δ in magnitude; thus one is led to consider the product

$$\prod_{n=p_1}^{p_1+p_2} |1 + \beta_n^{(m)}\nu_k| \,\delta \tag{40}$$

for all k, p_1, p_2. This product gives a more realistic estimate for the upper bound of $c_k^{(p_1+p_2)}$, and may become so large that the convergence of the method will be slowed or even prevented, and in severe cases the capacity of the machine will be exceeded.

The growth of roundoff errors may be inhibited by using the $\beta_n^{(m)}$ in a certain order within each cycle. It is clear that to use the $\beta_n^{(m)}$ in descending order of magnitude, as in (21), is not suitable since for large ν_k the largest of the factors $|1 + \beta_n^{(m)}\nu_k|$ in (40) are used first. It is somewhat better to use the order of ascending magnitudes. In this case the product $\prod_{n=1}^{p} |(1 + \beta_n^{(m)}\nu_k)|$ is reduced somewhat before the large factors are used. However the product cannot become smaller than δ because of roundoff. Then as the large $\beta_n^{(m)}$ are used the coefficient of $v^{(k)}$ may become as large as

$$\prod_n |1 + \beta_n^{(m)}\nu_k| \delta$$

where the product is taken over all factors such that $|1 + \beta_n^{(m)}\nu_k| \geqq 1$.

A better method, which has been more successful, is to use the ordering

$$\beta_{(m/2+1)}^{(m)}, \beta_{m/2}^{(m)}, \beta_{(m/2+2)}^{(m)}, \cdots, \beta_m^{(m)}, \beta_1^{(m)}$$

when m is even, and

$$\beta_{(m+1)/2}^{(m)}, \beta_{[(m+1)/2+1]}^{(m)}, \beta_{[(m+1)/2-1]}^{(m)}, \cdots, \beta_m^{(m)}, \beta_1^{(m)}$$

when m is odd. Here we assume that the $\beta_n^{(m)}$ are labeled with descending magnitude, i.e. $|\beta_1^{(m)}| \geqq |\beta_2^{(m)}| \geqq \cdots \geqq |\beta_m^{(m)}|$. This procedure has the advantage of using alternately large and small $\beta_n^{(m)}$, thus preventing uninterrupted growth of roundoff errors. Of course, still further improvements are possible.

We remark that the growth of roundoff errors can be considerably reduced if one uses polynomial operators, as was done by Shortley, [14]. However, as already noted, this requires the storage of many additional numbers.

5. Comparison with other methods. We now discuss briefly the relation of Richardson's method with other iterative methods such as the Gauss-Seidel method, the Successive Overrelaxation Method [11], and gradient methods including the method of steepest descent [10] and [4] and the method of conjugate gradients, [9] and [4].

With gradient methods one modifies the tentative solution at each stage in such a way as to reduce the value of a certain positive definite quadratic form. These methods appear to be very promising, especially the method of conjugate gradients where theoretically one obtains the exact solution after N iterations. Nevertheless, a considerable amount of further study and testing on computing machines remains before these methods can be accurately evaluated as tools for solving linear systems. Of course the same is true of Richardson's method, but it appears that the gradient methods require more computations per iteration. Furthermore, although estimates of the maximum and minimum eigenvalues are not needed to insure theoretical convergence for the gradient methods, never-

theless, these estimates may be necessary in actual practice in order to control the growth of roundoff errors.

In [11] it is shown that under certain assumptions on A, the Successive Overrelaxation Method (S.O.R. Method) defined by

$$(41)\qquad \begin{cases} u_i^{(n+1)} = -(\omega/a_{i,i})\left(\sum_{j=1}^{i-1} a_{i,j}u_j^{(n+1)} + \sum_{j=i+1}^{N} a_{i,j}u_j^{(n)} + d_i\right) \\ \qquad\qquad -(\omega-1)u_i^{(n)}, \qquad (i = 1, 2, \cdots, N) \\ u_i^{(0)} \text{ arbitrary}, \qquad (i = 1, 2, \cdots, N) \end{cases}$$

gives the same order-of-magnitude gain in convergence rate as Richardson's method, provided ω is suitably chosen. In fact, it can be shown that if $a_{i,i}$ is independent of i, then the rate of convergence of the method is given by

$$(42)\qquad R = -2\log\alpha$$

provided

$$(43)\qquad \omega = 1 + \alpha^2.$$

If in (41) we let $\omega = 1$ we obtain the Gauss-Seidel method. Under the same assumptions on A it is shown in [11] that the Gauss-Seidel method converges twice as fast as Richardson's method with $\beta_n = -2/(b + a)$ for all n, provided $a_{i,i}$ is independent of i. Thus Richardson's method with larger m is certainly superior to the Gauss-Seidel method.

The assumptions on A which were made to prove the preceding results are that A be symmetric and positive definite and also have Property (A) defined as follows:

There exist two disjoint subsets P and Q of W, the set of the first N positive integers, such that if $a_{i,j} \neq 0$, then $i = j$ or $i \in P$ and $j \in Q$ or $i \in Q$ and $j \in P$.

Evidently Richardson's method applies under more general conditions. Nevertheless, as shown in [11], for linear systems associated with many elliptic partial difference equations, the matrix does have Property (A). In these cases the S.O.R. Method has the following advantages for use on large automatic computing machines:

1. The storage problem is simpler since new iterated values are used as soon as obtained; hence the old values need not be retained until after a whole iteration has been completed, as is necessary with Richardson's method.

2. If the diagonal elements of A are equal, then the S.O.R. Method converges at least twice as fast as Richardson's method. This follows from (42) and (30). It appears likely that the S.O.R. Method is faster in all cases but this remains to be proved.

3. The S.O.R. Method is simpler since only one value of ω is needed while for Richardson's method many different β_n are used. Moreover, roundoff errors do not build up appreciably with the S.O.R. Method since $\omega \leq 2$. On the other hand roundoff is a very serious problem with Richardson's method.

BIBLIOGRAPHY

1. A. Brauer, *Limits for the characteristic roots of a matrix*, Duke Math. Jour., **13**, pp. 387–395 (1946).
2. D. Flanders and G. Shortley, *Numerical determination of fundamental modes*, J. App. Phys., **21**, pp. 1326–1332 (1950).
3. M. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, Journal of Research, **49**, pp. 409–436 (1952).
4. L. V. Kantorovich, *On an effective method of solving extremal problems for quadratic functionals*, Comptes Rendus (Doklady) de l'Académie des Sciences de L'URSS, 1945, **XLVIII**, Nov. 7.
5. W. Markoff, *Über Polynome, die in einem gegeben Intervalle möglichst wenig von Null abweichen*, Math. Ann., **77**, (1916), pp. 213–258. (Translation and condensation by J. Grossman of Russian article published in 1892.)
6. L. F. Richardson, *The approximate solution by finite differences of physical problems involving differential equations, with an application to the stresses in a masonry dam*, Roy. Soc. Phil. Trans. **210A**, pp. 307–357 (1910).
7. G. Shortley and R. Weller, *The numerical solution of Laplace's equation*, J. App. Phys., **9**, pp. 334–344 (1938).
8. R. Southwell, *Relaxation methods in theoretical physics*, University Press, Oxford, 1946.
9. E. Stiefel, *Über einige Methoden der Relaxationsrechnung*, Zeits. f. angew. Math. u. Phys., **3**, pp. 1–33 (1952).
10. G. Temple, *The general theory of relaxation methods applied to linear systems*, Proc. Roy. Soc., **A169**, pp. 476–500 (1938–39).
11. D. Young, *Iterative methods for solving partial difference equations of elliptic type*, doctoral dissertation, Harvard Univ. 1950. A revised version has been accepted for publication.
12. D. Young, *On Richardson's method for solving linear systems with positive definite matrices*, abstract of a paper presented at the October 25, 1952 meeting of the American Mathematical Society, New Haven, Conn. Abstract 26t, Bull. Amer. Math. Soc., **59**, pp. 47–48, 1953.
13. C. Lanczos, *Solution of systems of linear equations by minimized iterations*, Journal of Research, **49**, pp. 33–53 (1952).
14. G. Shortley, *Use of Tschebyscheff-polynomial operators in the numerical solution of boundary-value problems*, J. App. Phys., **24**, pp. 392–396 (1953).
15. D. Young and C. Warlick, *On the use of Richardson's method for the numerical solution of Laplace's equation on the ORDVAC*, Ballistic Research Laboratories Memorandum Report No. 707, Aberdeen Proving Ground, Md., July 1953.

University of Maryland

(Received April 1, 1953)

A SYSTEM OF NON-LINEAR PARTIAL DIFFERENTIAL EQUATIONS ARISING IN HEAT TRANSFER

By S. Katz and A. M. Peiser

1. Introduction. The differential equations which describe the transfer of heat or of mass from a fluid to a bed of particles through which the fluid is flowing may be written [2]*

$$f(u)\,\partial u/\partial x = u - v = g(v)\,\partial v/\partial y,$$
$$u(o^+, y) = \varphi(y), \qquad v(x, o^+) = \psi(x), \qquad x > 0, y > 0, \tag{1.1}$$

where x is distance along the bed and y is time. In these equations, $u = u(x, y)$ and $v = v(x, y)$ represent the temperature (in the case of heat transfer) or the concentration (in the case of mass transfer) in the fluid and the bed respectively. The functions f and g describe physical properties of the materials involved; heat capacities in the case of heat transfer and equilibrium adsorption isotherms in the case of mass transfer.

In certain ranges of temperature or concentration, satisfactory approximations may be obtained by supposing f and g to be constant. In this case, the solutions are well-known [3]. However, when this assumption cannot be made, as, for example, in the case of very low temperatures, solutions must be found by numerical methods. The differential equations (1.1) are replaced by the difference equations

$$f(U_{m,n})\,(U_{m+1,n} - U_{m,n})/h = U_{m,n} - V_{m,n}$$
$$= g(V_{m,n})\,(V_{m,n+1} - V_{m,n})/k, \tag{1.2}$$
$$U_{0,n} = \varphi(nk), \qquad V_{m,0} = \psi(mh), \qquad m, n = 0, 1, 2, \cdots,$$

in which the continuous variables x and y have been replaced by the discrete variables mh and nk respectively.

In this paper we are concerned with the existence and uniqueness of the solution of (1.1) and with the convergence of $U_{m,n}$, $V_{m,n}$ to this solution as the interval lengths h, k approach zero. We shall assume throughout that the boundary functions φ and ψ are continuous, and that the functions f and g are continuous and satisfy the inequalities

$$|f(t)| \geqq \alpha > 0, \qquad |g(t)| \geqq \alpha > 0, \qquad -\infty < t < \infty.$$

Under these assumptions we prove:

(A) There exists a unique pair of functions, $u(x, y)$ and $v(x, y)$, continuous in $x \geqq 0$, $y \geqq 0$, which satisfy the differential equations (1.1).

(B) If $m = [x/h]$ and $n = [y/k]$, then

$$\lim_{h,k\to 0} U_{m,n} = u(x, y),$$
$$\lim_{h,k\to 0} V_{m,n} = v(x, y); \tag{1.4}$$

* Numbers in brackets refer to the bibliography at the end of the paper.

that is, the solution of the difference equations approaches the solution of the differential equations as the "mesh diameter" approaches zero.

Consider the integral equations

$$\begin{aligned} \int_{\phi(y)}^{u(x,y)} f(s)\,ds &= \int_0^x [u(\xi, y) - v(\xi, y)]\,d\xi, \\ \int_{\psi(x)}^{v(x,y)} g(s)\,ds &= \int_0^y [u(x, \eta) - v(x, \eta)]\,d\eta, \qquad x \geqq 0, \quad y \geqq 0. \end{aligned} \tag{1.5}$$

With the assumption of continuity of f and g, it is clear that if u, v is a continuous solution of (1.5), then $\partial u/\partial x$ and $\partial v/\partial y$ exist and u, v satisfies (1.1). Conversely, if u, v is a continuous solution of (1.1), then $\partial u/\partial x$ and $\partial v/\partial y$ are continuous. Hence integration of (1.1) with a simple change of variables shows that u, v is a solution of (1.5). Accordingly, to prove (A), it is sufficient to prove that (1.5) has a unique continuous solution. This is proved in Section 2.

The convergence theorem (B) is proved in Section 3. It seems to be extremely difficult to get good error estimates for the differences in (1.4) in the non-linear case. With the additional assumption of a Lipschitz condition on f and g, the method of Section 3 can be made to yield error estimates which grow exponentially with x and y. In the linear case, however, where the solution is known, much better estimates may be obtained. In the important case $f(t) = -g(t) \equiv 1$, $\psi(x) \equiv 0$, $\varphi(y) \equiv 1$, we shall show in Section 4 that

$$\begin{aligned} u(x, y) - U_{m,n} &= \tfrac{1}{2}k\,K_1(y, x) + \tfrac{1}{2}(h + k)\,[yK_1(y, x) - xK_0(y, x)] \\ &\qquad + O[(h + k)^2], \\ v(x, y) - V_{m,n} &= \tfrac{1}{2}h\,K_1(x, y) + \tfrac{1}{2}(h + k)\,[xK_1(x, y) - yK_0(x, y)] \\ &\qquad + O[(h + k)^2], \end{aligned} \tag{1.6}$$

where

$$K_n(x, y) = (y/x)^{n/2}\,I_n(2\sqrt{xy})e^{-x-y}.$$

The function $K_n(x, y)$ decreases exponentially with increasing x and y except in the neighborhood of $x = y$ where it goes to zero like $x^{-1/2}$.

2. Existence of Solutions. This section presents a proof that the integral equations (1.5) have solutions, unique and continuous in $x \geqq 0$, $y \geqq 0$. It will be convenient, here and in the next sections, to have the following lemma.

LEMMA 1. Let $a_n(x, y)$ and $b_n(x, y)$, $n = 1, 2, \cdots$, be non-negative in the rectangle $0 \leqq x \leqq X$, $0 \leqq y \leqq Y$, and satisfy there the inequalities

$$\begin{aligned} \alpha a_{n+1}(x, y) &\leqq p + \int_0^x [a_n(\xi, y) + b_n(\xi, y)]\,d\xi, \\ \alpha b_{n+1}(x, y) &\leqq p + \int_0^y [a_n(x, \eta) + b_n(x, \eta)]\,d\eta, \qquad n = 1, 2, \cdots, \end{aligned}$$

where $\alpha > 0$, $p \geqq 0$. Then in this rectangle, both $a_{n+1}(x, y)$ and $b_{n+1}(x, y)$ are less than or equal to

$$pS(x, y) + \underset{\substack{0 \leq \xi \leq x \\ 0 \leq \eta \leq y}}{\text{l.u.b}} \{a_1(\xi, \eta) + b_1(\xi, \eta)\} \frac{[(x + y)/\alpha]^n}{([n/2]!)^2}, \qquad n = 1, 2, \cdots,$$

where

$$S(x, y) = (2/\alpha)\,[1 + (x + y)/\alpha]e^{[(x+y)/\alpha]^2}.$$

Proof: Let

$$A_n(x, y) = \underset{\substack{0 \leq \xi \leq x \\ 0 \leq \eta \leq y}}{\text{l.u.b.}}\ a_n(\xi, \eta),$$

$$B_n(x, y) = \underset{\substack{0 \leq \xi \leq x \\ 0 \leq \eta \leq y}}{\text{l.u.b.}}\ b_n(\xi, n), \qquad n = 1, 2, \cdots.$$

Then the integral inequalities in the hypothesis imply

$$\alpha a_{n+1}(\xi, \eta) \leqq p + \int_0^{\xi} \{A_n(s, \eta) + B_n(s, \eta)\}\, ds,$$

$$\alpha b_{n+1}(\xi, \eta) \leqq p + \int_0^{\eta} \{A_n(\xi, t) + B_n(\xi, t)\}\, dt.$$

Taking upper bounds over $0 \leqq \xi \leqq x$, $0 \leqq \eta \leqq y$, we get, on account of the monotonicity of A_n and B_n,

$$\alpha A_{n+1}(x, y) \leqq p + \int_0^{x} \{A_n(s, y) + B_n(s, y)\}\, ds,$$

$$\alpha B_{n+1}(x, y) \leqq p + \int_0^{y} \{A_n(x, t) + B_n(x, t)\}\, dt, \qquad n = 1, 2, \cdots.$$

With

$$\theta_i(x, y) = \frac{1}{\alpha^i} \sum_{j=0}^{i-1} \binom{i-1}{j} \frac{x^{i-j}}{(i-j)!} \frac{y^j}{j!},$$

so that

$$\int_0^{x} [\theta_i(s, y) + \theta_i(y, s)]\, ds = \alpha\theta_{i+1}(x, y),$$

it may readily be verified by induction that

$$\begin{aligned} A_{n+1}(x, y) &\leqq \frac{p}{\alpha}\left[1 + 2\sum_{i=1}^{n-1} \theta_i(x, y)\right] + [A_1(x, y) + B_1(x, y)]\theta_n(x, y), \\ B_{n+1}(x, y) &\leqq \frac{p}{\alpha}\left[1 + 2\sum_{i=1}^{n-1} \theta_i(y, x)\right] + [A_1(x, y) + B_1(x, y)]\theta_n(y, x), \\ &\qquad n = 1, 2, \cdots. \end{aligned} \tag{2.1}$$

Now

$$j!(i-j)! \geqq \left(\left[\frac{i}{2}\right]!\right)^2, \qquad j = 0, \cdots, i,$$

so that

$$\begin{aligned} \theta_i(x, y) &\leqq \frac{1}{([i/2]!)^2}\left(\frac{x}{\alpha}\right)\sum_{j=0}^{i-1}\binom{i-1}{j}\left(\frac{x}{\alpha}\right)^{i-1-j}\left(\frac{y}{\alpha}\right)^{j} \\ &\leqq \frac{[(x+y)/\alpha]^i}{([i/2]!)^2}, \end{aligned} \tag{2.2}$$

and

$$\begin{aligned} 1 + \sum_{i=1}^{n-1}\theta_i(x, y) &\leqq \sum_{i=0}^{\infty}\frac{[(x+y)/\alpha]^i}{([i/2]!)^2} \\ &= \left(1 + \frac{x+y}{\alpha}\right)\sum_{k=0}^{\infty}\frac{[(x+y)/\alpha]^{2k}}{(k!)^2} \\ &\leqq \left(1 + \frac{x+y}{\alpha}\right)e^{[(x+y)/\alpha]^2}. \end{aligned} \tag{2.3}$$

Equations (2.1), (2.2), (2.3) imply the lemma.

A formal iterative procedure for solving the integral equations (15.) in $x \geqq 0$, $y \geqq 0$ may be set up as

$$\begin{aligned} u_0(x, y) &= 0, \qquad v_0(x, y) = 0, \\ \int_{\varphi(y)}^{u_{n+1}(x,y)} f(s)\,ds &= \int_0^x \{u_n(\xi, y) - v_n(\xi, y)\}\,d\xi \\ \int_{\psi(x)}^{v_{n+1}(x,y)} g(t)\,dt &= \int_0^y \{u_n(x, \eta) - v_n(x, \eta)\}\,d\eta, \qquad n = 0, 1, \cdots. \end{aligned} \tag{2.4}$$

That (2.4) really defines an iterative process is guaranteed by the fact that each of the functions f and g is of one sign and bounded away from 0. Thus each of the integrals

$$\int_{\varphi}^{u} f(s)\,ds, \qquad \int_{\psi}^{v} g(t)\,dt,$$

is either a strictly increasing or a strictly decreasing function of its upper limit, between $+\infty$ at one end of the real axis and $-\infty$ at the other. It may be noted further that since $f, g, \varphi, \psi, u_0, v_0$ are all continuous functions of their arguments in the appropriate ranges, successive convergents $u_n(x, y)$, $v_n(x, y)$ defined by (2.4) are continuous as well.

To establish the convergence of the iterative process (2.4), consider the differences between these equations for successive values of n:

$$\begin{aligned} \int_{u_n(x,y)}^{u_{n+1}(x,y)} f(s)\,ds &= \int_0^x [\{u_n(\xi, y) - u_{n-1}(\xi, y)\} - \{v_n(\xi, y) - v_{n-1}(\xi, y)\}]\,d\xi, \\ \int_{v_n(x,y)}^{v_{n+1}(x,y)} g(t)\,dt &= \int_0^y [\{u_n(x, \eta) - u_{n-1}(x, \eta)\} - \{v_n(x, \eta) - v_{n-1}(x, \eta)\}]\,d\eta, \\ & \qquad n = 1, 2, \cdots. \end{aligned} \tag{2.5}$$

With

$$a_n(x, y) = |\, u_n(x, y) - u_{n-1}(x, y)\,|, \qquad b_n(x, y) = |\, v_n(x, y) - v_{n-1}(x, y)\,|,$$

these equations, together with (1.3), give

$$\begin{aligned} \alpha a_{n+1}(x, y) &\leqq \int_0^x \{a_n(\xi, y) + b_n(\xi, y)\}\, d\xi, \\ \alpha b_{n+1}(x, y) &\leqq \int_0^y \{a_n(x, \eta) + b_n(x, \eta)\}\, d\eta, \qquad n = 1, 2, \cdots. \end{aligned} \tag{2.6}$$

Lemma 1, with $p = 0$, then states that both $a_{n+1}(x, y)$ and $b_{n+1}(x, y)$ are dominated by

$$\underset{\substack{0 \leqq \xi \leqq x \\ 0 \leqq \eta \leqq y}}{\text{l.u.b.}} \{\, |\varphi(\eta)| + |\psi(\xi)|\,\} \frac{[(x + y)/\alpha]^n}{([n/2]!)^2}, \qquad n = 1, 2, \cdots,$$

since

$$a_1(x, y) = u_1(x, y) = \varphi(y), \qquad b_1(x, y) = v_1(x, y) = \psi(x).$$

Hence the series

$$\begin{aligned} u(\xi, \eta) &= \textstyle\sum_{n=0}^{\infty} \{u_{n+1}(\xi, \eta) - u_n(\xi, \eta)\}, \\ v(\xi, \eta) &= \textstyle\sum_{n=0}^{\infty} \{v_{n+1}(\xi, \eta) - v_n(\xi, \eta)\} \end{aligned}$$

converge uniformly in every interval $0 \leqq \xi \leqq x$, $0 \leqq \eta \leqq y$. The functions u and v accordingly exist and are continuous for every $x \geqq 0$, $y \geqq 0$. Since the uniform convergence permits the passage to the limit under the integral sign in (2.4), u and v satisfy the integral equations (1.5).

To show that the solution of (1.5) is unique, consider two solutions: u, v and u', v'. The difference between the equations (1.5) for these two solutions is, as in (2.5),

$$\int_{u(x,y)}^{u'(x,y)} f(s)\, ds = \int_0^x [\{u'(\xi, y) - u(\xi, y)\} - \{v'(\xi, y) - v(\xi, y)\}]\, d\xi,$$

$$\int_{v(x,y)}^{v'(x,y)} g(t)\, dt = \int_0^y [\{u'(x, \eta) - u(x, \eta)\} - \{v'(x, \eta) - v(x, \eta)\}]\, d\eta.$$

Taking absolute values gives, as in (2.6), the inequalities

$$\begin{aligned} \alpha a(x, y) &\leqq \int_0^x \{a(\xi, y) + b(\xi, y)\}\, d\xi, \\ \alpha b(x, y) &\leqq \int_0^y \{a(x, \eta) + b(x, \eta)\}\, d\eta, \end{aligned}$$

where

$$a(x, y) = |\, u'(x, y) - u(x, y)\,|, \qquad b(x, y) = |\, v'(x, y) - v(x, y)\,|.$$

Lemma 1, with $p = 0$ and

$$a_n(x, y) = a(x, y), \qquad b_n(x, y) = b(x, y), \qquad n = 1, 2, \cdots,$$

then states that both $a(x, y)$ and $b(x, y)$ are, for every n, less than or equal to

$$\underset{\substack{0 \le \xi \le x \\ 0 \le \eta \le y}}{\text{l.u.b.}} \{a(\xi, \eta) + b(\xi, \eta)\} \frac{[(x + y)/\alpha]^n}{([n/2]!)^2}, \qquad n = 1, 2, \cdots.$$

Hence

$$a(x, y) = b(x, y) = 0,$$

and uniqueness is proved.

3. Convergence of $U_{m,n}$ and $V_{m,n}$. This section presents a proof that the solution $U_{m,n}$, $V_{m,n}$ of the difference equations (1.2) approaches the solution $u(x, y)$, $v(x, y)$ of the differential equations (1.1) as $h, k \to 0$. Fix a rectangle $0 \le x \le X$, $0 \le y \le Y$, and suppose throughout this section that all values of the continuous variables x, y and of the discrete variables mh, nk lie within this rectangle. We require first a bound on $U_{m,n} - V_{m,n}$.

LEMMA 2. For all h and k,

$$|U_{m,n} - V_{m,n}| \le R(X, Y) = \underset{\substack{0 \le x \le X \\ 0 \le y \le Y}}{\text{l.u.b.}} \{|\varphi(y)| + |\psi(x)|\} e^{(2/\alpha)(X+Y)}.$$

Proof:

We shall prove the stronger inequality,

$$\Delta_{m,n} = |U_{m,n} - V_{m,n}| \le \{\Phi(nk) + \Psi(mh)\} e^{(2/\alpha)(mh+nk)}, \tag{3.1}$$

where

$$\Phi(nk) = \max_{0 \le j \le n} |\varphi(jk)|, \qquad \Psi(mh) = \max_{0 \le i \le m} |\psi(ih)|.$$

Now (1.2) may be summed to give

$$U_{m,n} = \varphi(nk) + \sum_{i=0}^{m-1} \frac{h}{f(U_{i,n})} (U_{i,n} - V_{i,n}),$$

$$V_{m,n} = \psi(mh) + \sum_{j=0}^{n-1} \frac{k}{g(V_{m,j})} (U_{m,j} - V_{m,j}).$$

Subtracting the second equation from the first, it is easy to see that

$$\Delta_{m,n} \le \Phi(nk) + \Psi(mh) + (h/\alpha) \textstyle\sum_{i=0}^{m-1} \Delta_{i,n} + (k/\alpha) \sum_{j=0}^{n-1} \Delta_{m,j}. \tag{3.2}$$

We prove (3.1) by induction. In the case $n = 0$, (3.2) reduces to

$$\Delta_{m,0} \le \Phi(0) + \Psi(mh) + (h/\alpha) \textstyle\sum_{i=0}^{m-1} \Delta_{i,0}.$$

That $\Delta_{m,0}$ satisfies the inequality (3.1) may be seen by an induction on m. Clearly (3.1) holds for $\Delta_{0,0}$ and if it holds for $\Delta_{i,0}$, $0 \le i \le m - 1$, then

$$\Delta_{m,0} \leqq \{\Phi(0) + \Psi(mh)\}\left[1 + (h/\alpha)\sum_{i=0}^{m-1} e^{(2/\alpha)ih}\right]$$
$$= \{\Phi(0) + \Psi(mh)\}\left[1 + (h/\alpha)\frac{e^{(2/\alpha)mh} - 1}{e^{(2/\alpha)h} - 1}\right].$$

With the inequality

$$w/(e^w - 1) < 1, \qquad w > 0. \tag{3.3}$$

we have

$$\Delta_{m,0} \leqq \{\Phi(0) + \Psi(mh)\}\,[1 + \tfrac{1}{2}(e^{(2/\alpha)mh} - 1)] \leqq \{\Phi(0) + \Psi(mh)\}e^{(2/\alpha)mh}.$$

Similarly, it may be shown that $\Delta_{0,n}$ satisfies (3.1).

Suppose now that (3.1) holds for $i = m$, $0 \leqq j \leqq n - 1$, and for $j = n$, $0 \leqq i \leqq m - 1$.
Then (3.2) yields

$$\Delta_{m,n} \leqq \{\Phi(nk) + \Psi(mh)\}\left[1 + \frac{h}{\alpha}\sum_{i=0}^{m-1} e^{(2/\alpha)(ih+nk)} + \frac{k}{\alpha}\sum_{j=0}^{n-1} e^{(2/\alpha)(mh+jk)}\right]$$
$$= \{\Phi(nk) + \Psi(mh)\}$$
$$\cdot\left[1 + \frac{h}{\alpha}e^{(2/\alpha)nk}\frac{e^{(2/\alpha)mh} - 1}{e^{(2/\alpha)h} - 1} + \frac{k}{\alpha}e^{(2/\alpha)mh}\frac{e^{(2/\alpha)nk} - 1}{e^{(2/\alpha)k} - 1}\right].$$

With the inequality (3.3), we obtain

$$\Delta_{m,n} \leqq \{\Phi(nk) + \Psi(mh)\}$$
$$\cdot[1 + \tfrac{1}{2}e^{(2/\alpha)nk}(e^{(2/\alpha)mh} - 1) + \tfrac{1}{2}e^{(2/\alpha)mh}(e^{(2/\alpha)nk} - 1)]$$
$$\leqq \{\Phi(nk) + \Psi(mh)\}e^{(2/\alpha)(mh+nk)}.$$

This proves the lemma.

Lemma 2, together with the difference equations (1.2), have two immediate consequences: the boundedness of the difference quotients of $U_{m,n}$ and $V_{m,n}$,

$$|U_{m+1,n} - U_{m,n}| \leqq (h/\alpha)R(X, Y), \qquad |V_{m,n+1} - V_{m,n}| \leqq (k/\alpha)R(X, Y); \tag{3.4}$$

and, as a result, the boundedness of $U_{m,n}$ and $V_{m,n}$ for all h and k,

$$|U_{m,n}| \leqq \underset{0\leqq y\leqq Y}{\text{l.u.b.}}\,|\varphi(y)| + (1/\alpha)XR(X, Y),$$
$$|V_{m,n}| \leqq \underset{0\leqq x\leqq X}{\text{l.u.b.}}\,|\psi(x)| + (1/\alpha)YR(X, Y). \tag{3.5}$$

Now if we write, for a fixed h and k,

$$\left.\begin{aligned} U(x, y) &= U_{m,n} \\ V(x, y) &= V_{m,n} \end{aligned}\right\} \begin{aligned} mh &\leqq x < (m + 1)h \\ nh &\leqq y < (n + 1)k, \end{aligned} \qquad m, n = 0, 1, \cdots, \tag{3.6}$$

the difference equations (1.2) sum to

$$\sum_{i=0}^{m-1} f(U_{i,n})(U_{i+1,n} - U_{i,n}) = \int_0^{mh} [U(\xi, nk) - V(\xi, nk)]\, d\xi,$$
$$\sum_{j=0}^{n-1} g(V_{m,j})(V_{m,j+1} - V_{m,j}) = \int_0^{nk} [U(mh, \eta) - V(mh, \eta)]\, d\eta. \tag{3.7}$$

The integrals in the left members of (1.5) can be broken up into sums like those in (3.7), giving, for $x = mh$, $y = nk$,

$$\int_{U(mh,nk)}^{u(mh,nk)} f(s)\, ds + \sum_{i=0}^{m-1} \int_{U_{i,n}}^{U_{i+1,n}} f(s)\, ds = \int_0^{mh} [u(\xi, nk) - v(\xi, nk)]\, d\xi,$$
$$\int_{V(mh,nk)}^{v(mh,nk)} g(t)\, dt + \sum_{j=0}^{n-1} \int_{V_{m,j}}^{V_{m,j+1}} g(t)\, dt = \int_0^{nk} [u(mh, \eta) - v(mh, \eta)]\, d\eta. \tag{3.8}$$

Subtracting (3.7) from (3.8) gives

$$\int_{U(mh,nk)}^{u(mh,nk)} f(s)\, ds = \int_0^{mh} [P(\xi, nk) - Q(\xi, nk)]\, d\xi - \sum_{i=0}^{m-1} \int_{U_{i,n}}^{U_{i+1,n}} \{f(s) - f(U_{i,n})\}\, ds,$$
$$\int_{V(mh,nk)}^{v(mh,nk)} g(t)\, dt = \int_0^{nk} [P(mh, \eta) - Q(mh, \eta)]\, d\eta - \sum_{j=0}^{n-1} \int_{V_{m,j}}^{V_{m,j+1}} \{g(t) - g(V_{m,j})\}\, dt, \tag{3.9}$$

where

$$P(x, y) = u(x, y) - U(x, y), \qquad Q(x, y) = v(x, y) - V(x, y). \tag{3.10}$$

Now f and g are uniformly continuous in the closed intervals (3.5). Hence, corresponding to an arbitrary $\epsilon > 0$, we can select h and k small enough to ensure that

$$|f(s) - f(\sigma)| \leqq \frac{\alpha\epsilon}{4XR(X, Y)S(X, Y)}, \qquad |s - \sigma| \leqq \frac{h}{\alpha} R(X, Y),$$
$$|g(t) - g(\tau)| \leqq \frac{\alpha\epsilon}{4YR(X, Y)S(X, Y)}, \qquad |t - \tau| \leqq \frac{k}{\alpha} R(X, Y), \tag{3.11}$$

with the R of Lemma 2 and the S of Lemma 1. The inequalities (3.4) then give

$$\left| \sum_{i=0}^{m-1} \int_{U_{i,n}}^{U_{i+1,n}} \{f(s) - f(U_{i,n})\}\, ds \right| \leqq m \cdot \frac{h}{\alpha} R(X, Y) \cdot \frac{\alpha\epsilon}{4XR(X, Y)S(X, Y)} \leqq \frac{\epsilon}{4S(X, Y)},$$
$$\left| \sum_{j=0}^{n-1} \int_{V_{m,j}}^{V_{m,j+1}} \{g(t) - g(V_{m,j})\}\, dt \right| \leqq n \cdot \frac{k}{\alpha} R(X, Y) \cdot \frac{\alpha\epsilon}{4YR(X, Y)S(X, Y)} \leqq \frac{\epsilon}{4S(X, Y)}. \tag{3.12}$$

Also, the lower bounds (1.3) on $|f|$ and $|g|$ imply that

$$\left| \int_{U(mh,nk)}^{u(mh,nk)} f(s)\, ds \right| \geqq \alpha \mid P(mh, nk) \mid ,$$
$$\left| \int_{V(mh,nk)}^{v(mh,nk)} g(t)\, dt \right| \geqq \alpha \mid Q(mh, nk) \mid . \tag{3.13}$$

Applying (3.12) and (3.13) to (3.9) gives

$$\alpha \mid P(mh, nk) \mid \leqq \frac{\epsilon}{4S(X, Y)} + \int_0^{mh} [\mid P(\xi, nk) \mid + \mid Q(\xi, nk) \mid] \, d\xi,$$
$$\alpha \mid Q(mh, nk) \mid \leqq \frac{\epsilon}{4S(X, Y)} + \int_0^{nk} [\mid P(vh, \eta) \mid + \mid Q(mh, \eta) \mid] \, d\eta. \tag{3.14}$$

To rewrite (3.14) in terms of the continuous variables x, y, we note that $u(x, y)$ and $v(x, y)$ are uniformly continuous in the rectangle $0 \leqq x \leqq X, 0 \leqq y \leqq Y$. Hence, for the arbitrary positive ϵ of (3.11), h and k can be chosen small enough to satisfy

$$\left.\begin{aligned} \mid u(x, y) - u(\xi, \eta) \mid &\leqq \frac{\epsilon}{4S(X, Y)} \min. \left(\frac{1}{\alpha}, \frac{1}{X}, \frac{1}{Y}\right), \\ \mid v(x, y) - v(\xi, \eta) \mid &\leqq \frac{\epsilon}{4S(X, Y)} \min. \left(\frac{1}{\alpha}, \frac{1}{X}, \frac{1}{Y}\right) \end{aligned}\right\} \quad \begin{aligned} &\mid x - \xi \mid \leqq h, \\ &\mid y - \eta \mid \leqq k \end{aligned} \tag{3.15}$$

as well as (3.11). By (3.6) and (3.10), the variation of P or Q over a rectangle of the mesh is simply the variation of u or v over the rectangle. Thus (3.14) with $m = [x/h]$ and $n = [y/k]$, together with (3.15), imply

$$\alpha \mid P(x, y) \mid \leqq \frac{\epsilon}{S(X, Y)} + \int_0^x [\mid P(\xi, y) \mid + \mid Q(\xi, y) \mid] \, d\xi,$$
$$\alpha \mid Q(x, y) \mid \leqq \frac{\epsilon}{S(X, Y)} + \int_0^y [\mid P(x, \eta) \mid + \mid Q(x, \eta) \mid] \, d\eta. \tag{3.16}$$

Applying Lemma 1 to (3.16), with

$$p = \frac{\epsilon}{S(X, Y)}, \quad a_n(x, y) = \mid P(x, y) \mid , \quad b_n(x, y) = \mid Q(x, y) \mid , \quad n = 1, 2, \cdots ,$$

gives, on account of the monotonicity of S,

$$\mid P(x, y) \mid \leqq \epsilon + \underset{\substack{0 \leqq \xi \leqq x \\ 0 \leqq \eta \leqq y}}{\text{l.u.b.}} \{\mid P(\xi, \eta) \mid + \mid Q(\xi, \eta) \mid\} \frac{[(x + y)/\alpha]^n}{([n/2]!)^2},$$

$$\mid Q(x, y) \mid \leqq \epsilon + \underset{\substack{0 \leqq \xi \leqq x \\ 0 \leqq \eta \leqq y}}{\text{l.u.b.}} \{\mid P(\xi, \eta) \mid + \mid Q(\xi, \eta) \mid\} \frac{[(x + y)/\alpha]^n}{([n/2]!)^2}, \quad n = 1, 2, \cdots .$$

Hence

$$\mid P(x, y) \mid \leqq \epsilon, \qquad \mid Q(x, y) \mid \leqq \epsilon,$$

which completes the proof.

4. Error Estimates in the Linear Case. A case of considerable practical importance occurs when $f(x)$ and $g(x)$ are taken as constants of *opposite* sign and the boundary values are also assumed to be constant. In this case explicit solutions may be found for both the differential and difference equations, and we shall obtain bounds for the difference between them. There is no loss of generality in taking the equations (1.5) in the form

$$-\partial u/\partial x = u - v = \partial v/\partial y, \qquad u(0, y) = 1, \qquad v(x, 0) = 0. \tag{4.1}$$

The solution in this case is known to be [3]

$$\begin{aligned} u(x, y) &= 1 - \int_0^x K_0(s, y)\, ds, \\ v(x, y) &= \int_0^y K_0(s, x)\, ds, \end{aligned} \tag{4.2}$$

where

$$K_n(x, y) = (y/x)^{(n/2)} e^{-x-y} I_n(2\sqrt{xy}), \qquad n = 0, 1, 2, \cdots. \tag{4.3}$$

The corresponding difference equations are

$$\frac{1}{h}(U_{m,n} - U_{m+1,n}) = U_{m,n} - V_{m,n} = \frac{1}{k}(V_{m,n+1} - V_{m,n}).$$

$$U_{0,n} = 1, \qquad V_{m,0} = 0,$$

and it is easy to see, upon elimination, that both $U_{m,n}$ and $V_{m,n}$ satisfy the equation

$$T_{m+1,n+1} = (1 - h)T_{m,n+1} + (1 - k)T_{m+1,n} - (1 - h - k)T_{m,n}, \tag{4.4}$$

with the boundary conditions

$$U_{0,n} = 1, \qquad U_{m,0} = (1 - h)^m, \tag{4.5}$$

and

$$V_{0,n} = 1 - (1 - k)^n, \qquad V_{m,0} = 0, \tag{4.6}$$

respectively. We suppose $0 < h < 1$, $0 < k < 1$.

Equation (4.4) may be solved by the method of separation of variables. Writing $T = w^m z^n$, we find

$$wz = (1 - h)z + (1 - k)w - (1 - h - k),$$

so that

$$z = 1 - k - [hk/(1 - h - w)].$$

Hence, for any appropriate function $f(w)$ of the complex variable w, the integral

$$\frac{1}{2\pi i}\int_{C'} w^m \left(1 - k - \frac{hk}{1 - h - w}\right)^n f(w)\, dw$$

is a solution of (4.4). The contour C' and the function f must be chosen so as to satisfy the conditions (4.5) or (4.6). To satisfy (4.6), let C' encircle the point $w = 1 - h$. Putting $n = 0$, we see that $f(w)$ may be taken to be analytic at $w = 1 - h$ with C' chosen so as to exclude the singularities of $f(w)$. Thus, we can write

$$f(w) = \sum_{j=0}^{\infty} p_j(1 - h - w)^j,$$

and, putting $m = 0$, the p_j must be chosen so that the residue of the resulting integrand at $w = 1 - h$ is $[1 - (1 - k)^n]$. But it is clear that this residue is given by the coefficient of $(w - 1 + h)^{-1}$ in the sum

$$\sum_{i=0}^{n} \binom{n}{i} (1 - k)^{n-i}(-hk)^i \sum_{j=0}^{\infty} p_j(1 - h - w)^{j-i},$$

so that

$$\sum_{i=1}^{n} \binom{n}{i} (1 - k)^{n-i}(-1)^{i-1}(hk)^i p_{i-1} = 1 - (1 - k)^n.$$

This equation is satisfied with $p_j = (-1)^j/h^{j+1}$, so that

$$f(w) = \sum_{j=0}^{\infty} (-1)^j \frac{(1 - h - w)^j}{h^{j+1}} = \frac{1}{1 - w}.$$

Hence

$$V_{m,n} = \frac{1}{2\pi i} \int_{C'} w^m \left(1 - k - \frac{hk}{1 - h - w}\right)^n \frac{dw}{1 - w}, \tag{4.7}$$

where C' includes the point $w = 1 - h$, and excludes the point $w = 1$.

As may be seen from a comparison of the boundary conditions (4.5) and (4.6), an expression for $U_{m,n}$ may be obtained directly from (4.7) by replacing V by $1 - U$, interchanging m and n, and interchanging h and k:

$$U_{m,n} = 1 - \frac{1}{2\pi i} \int_{C'} w^n \left(1 - h - \frac{hk}{1 - k - w}\right)^m \frac{dw}{1 - w}. \tag{4.8}$$

The contour C' must now also include the point $w = 1 - k$.

To facilitate a comparison between $v(x, y)$ and $V_{m,n}$, for example, it is convenient to express $v(x, y)$ as a contour integral. This may be accomplished by means of the Laplace transform. By standard integration formulas [1],

$$\int_0^{\infty} e^{-xz} K_0(y, x)\, dx = \frac{1}{z + 1} \exp\left(\frac{-yz}{z + 1}\right), \qquad Rz > -1,$$

so that, from (4.2),

$$\int_0^{\infty} e^{-xz} v(x, y)\, dx = \frac{1}{z}\left[1 - \exp\left(\frac{-yz}{z + 1}\right)\right], \qquad Rz > -1.$$

By the complex inversion formula [5],

$$v(x, y) = \frac{1}{2\pi i} \int_{\Gamma} \exp(xz)\left[1 - \exp\left(\frac{-yz}{z + 1}\right)\right] \frac{dz}{z},$$

where Γ is any vertical line $Rz = \lambda > -1$, and where the principal value of the integral is to be taken. By an application of Cauchy's theorem, Γ may be replaced by a closed contour C encircling the point $z = -1$. If C excludes the origin, then

$$v(x, y) = -\frac{1}{2\pi i}\int_C \exp\left(xz - \frac{yz}{z+1}\right)\frac{dz}{z}. \tag{4.9}$$

We suppose now that $x = mh$, $y = nk$, with $0 < h < 1$, $0 < k < 1$. With the change of variables $w = 1 + hz$, (4.7) becomes

$$V_{m,n} = -\frac{1}{2\pi i}\int_C (1+hz)^{x/h}\left(1 - \frac{kz}{z+1}\right)^{y/k}\frac{dz}{z}. \tag{4.10}$$

The contour C, which includes the point $z = -1$ but excludes the point $z = 0$, may be taken the same in (4.9) as in (4.10). Now,

$$\begin{aligned}(1+hz)^{x/h} &= \exp\left[\frac{x}{h}\log(1+hz)\right] \\ &= \exp\left[\frac{x}{h}\left\{hz - \frac{h^2z^2}{2} + O(h^3)\right\}\right] \\ &= \exp\left[xz - \frac{hxz^2}{2} + O(h^2)\right] \\ &= \exp(xz)\left[1 - \frac{hxz^2}{2} + O(h^2)\right],\end{aligned}$$

and applying this and a similar result to (4.10), we find

$$V_{m,n} = -\frac{1}{2\pi i}\int_C \exp\left(xz - \frac{yz}{z+1}\right) \cdot\left[1 - \frac{hx}{2}z^2 - \frac{ky}{2}\left(\frac{z}{z+1}\right)^2 + O\{(h+k)^2\}\right]\frac{dz}{z}.$$

Thus, from (4.9), with $x = mh$, $y = nk$,

$$\begin{aligned}v(x, y) - V_{m,n} &= \frac{hx}{2}\frac{1}{2\pi i}\int_C z\exp\left(xz - \frac{yz}{z+1}\right)dz \\ &\quad + \frac{ky}{2}\frac{1}{2\pi i}\int_C \frac{z}{(z+1)^2}\exp\left(xz - \frac{yz}{z+1}\right)dz + O\{(h+k)^2\} \qquad (4.11) \\ &= \frac{hx}{2}v_{xx}(x, y) + \frac{ky}{2}v_{yy}(x, y) + O\{(h+k)^2\}.\end{aligned}$$

Returning to (4.2) and (4.3), the elementary properties of the Bessel functions [4] yield

$$v_x(x, y) = \int_0^y \frac{\partial K_0(s, x)}{\partial x}\,ds = -K_1(x, y),$$

$$v_{xx}(x, y) = -\frac{\partial K_1(x, y)}{\partial x} = \left(1 + \frac{1}{x}\right)K_1(x, y) - \frac{y}{x}K_0(x, y),$$

and

$$v_{xy}(x, y) = \frac{\partial K_0(x, y)}{\partial y} = \frac{x}{y} K_1(x, y) - K_0(x, y).$$

Substitution of these results in (4.11) yields (1.6). The result for $u(x, y) - U_{m,n}$ is obtained in a similar way.

BIBLIOGRAPHY

(1) R. V. Churchill, "Modern Operational Mathematics in Engineering," New York, 1944, pp. 294, 299.
(2) E. Ledoux, "Dynamic Cooling of Adsorbent Beds," Industrial and Engineering Chemistry, **40**, (1948), pp. 1970–1977.
(3) A. M. Peiser, "Uniform Approximations to a Class of Bessel Functions," Proc. Am. Math. Soc., **1**, (1950), pp. 650–661.
(4) G. N. Watson, "Bessel Functions," Cambridge, 1945, p. 79.
(5) D. V. Widder, "The Laplace Transform," Princeton, 1946, p. 241.

Hydrocarbon Research, Inc.
New York, N. Y.

(Received July 22, 1952)

RECURRENCE RELATIONS FOR PROLATE SPHEROIDAL WAVE FUNCTIONS

IMANUEL MARX

1. Introduction. Recent investigations of electromagnetic problems in regions bounded by spheroids have led to an increased interest in the spheroidal wave functions. These functions $S_{mn}(c, \eta)$, $R_{mn}(c, \xi)$ result from the solution of the scalar wave equation by separation in spheroidal coordinates. The properties of S_{mn} and R_{mn} are discussed in [1] and [6]. The notation throughout this paper will follow that of [1].

Recurrence relations among the spheroidal wave functions cannot be established by the methods for functions of hypergeometric type, because the differential equation satisfied by these functions has an irregular singular point at infinity in addition to regular singular points for the values ± 1 of the respective variable. However, a method suggested by E. T. Whittaker [7], but apparently not fully exploited by him, leads to recurrence relations for the radial functions $R_{mn}(c, \xi)$ as well as for the non-integral separation constants that appear in the differential equations of both the radial and the angular functions. Formulas are obtained in detail for the prolate spheroidal wave functions. A simple substitution [4, p. 151] formally gives the corresponding relations for the oblate spheroidal functions, and the validity of these relations can be established by a suitable modification of the proof.

Whittaker observes that in the case of the two-dimensional wave equation $\partial^2 u/\partial x^2 + \partial^2 u/\partial y^2 + u = 0$ the derivative $\partial u/\partial x$ of a solution $u(x, y)$ is likewise a solution. After separation of the equation in polar coordinates, $\partial u/\partial x$ can be expanded in a series of separated solutions $J_n(r) \cos n\theta$. Write $\partial u/\partial x = \cos\theta\, \partial u/\partial r - r^{-1} \sin\theta\, \partial u/\partial\theta$, and equate to the series of Bessel functions. The resulting identity, when multiplied by $\cos p\theta$ and integrated from $-\pi$ to π, yields recurrence relations for the Bessel functions in terms of the coefficients in the expansion of $\partial u/\partial x$. Whittaker does not show how to obtain these coefficients. In the present three dimensional case, the analogous coefficients are represented as integrals of products of the spheroidal wave functions. Since these integrals involve the functions whose recurrence relations are being investigated, the existence of these relations and their nature, rather than a practically applicable form, are demonstrated. For practical applications, an evaluation of the coefficients by some independent method (such as is furnished for many known functions by the hypergeometric series) would seem to be necessary. It is hoped that further investigation will lead to such a method.

Procedures similar to those used in this paper may lead to recurrence relations in other cases of separable partial differential equations with constant coefficients. As long as the separated equations are of Sturm-Liouville type, and the solutions can be shown to have suitable orthogonality properties, an analogous method may be developed.

2. Properties of Wave Functions. The solutions $\Psi(x, y, z)$ of the scalar wave equation

$$\frac{\partial^2\Psi}{\partial x^2} + \frac{\partial^2\Psi}{\partial y^2} + \frac{\partial^2\Psi}{\partial z^2} + k^2\Psi = 0 \tag{1}$$

are to be obtained by a separation of the form

$$\Psi(x, y, z) = \Omega(\eta, \xi, \phi) = S(\eta)R(\xi)\Phi(\phi),$$

where

$$x = \tfrac{1}{2}d(1 - \eta^2)^{\frac{1}{2}}(\xi^2 - 1)^{\frac{1}{2}} \cos\phi,$$

$$y = \tfrac{1}{2}d(1 - \eta^2)^{\frac{1}{2}}(\xi^2 - 1)^{\frac{1}{2}} \sin\phi,$$

$$z = \tfrac{1}{2}d\eta\xi.$$

The surfaces $\eta = \text{const.}$, $-1 \leqq \eta \leqq 1$, are confocal hyperboloids of revolution of two sheets, the surfaces $\xi = \text{const.}$, $1 \leqq \xi < \infty$, are confocal prolate spheroids, and d is the focal distance common to both sets of surfaces. The surfaces $\phi = \text{const.}$, $0 \leqq \phi < 2\pi$, are half-planes containing the z-axis and terminated by it.

In terms of the spheroidal coordinates, equation (1) becomes

$$\frac{\partial}{\partial\eta}(1 - \eta^2)\frac{\partial\psi}{\partial\eta} + \frac{\partial}{\partial\xi}(\xi^2 - 1)\frac{\partial\psi}{\partial\xi} + \frac{(\xi^2 - \eta^2)}{(\xi^2 - 1)(1 - \eta^2)}\frac{\partial^2\psi}{\partial\varphi^2} + c^2(\xi^2 - \eta^2)\psi = 0,$$

where $c = \frac{1}{2}kd$. The separation of $\Phi(\phi)$ leads to the equation $\Phi'' + m^2\Phi = 0$. The separation constant m must be integral or zero to satisfy the condition of periodicity. Further separation gives a separation constant λ_{mn} and the differential equations

$$\frac{d}{d\eta}(1 - \eta^2)\frac{dS_{mn}}{d\eta} - \left(\frac{m^2}{1 - \eta^2} - \lambda_{mn} + c^2\eta^2\right)S_{mn} = 0, \tag{2}$$

$$\frac{d}{d\xi}(\xi^2 - 1)\frac{dR_{mn}}{d\xi} - \left(\frac{m^2}{\xi^2 - 1} + \lambda_{mn} - c^2\xi^2\right)R_{mn} = 0. \tag{3}$$

The separation constant is chosen so as to ensure that the solution of equation (2) which is free of logarithmic terms is regular in the entire range $-1 \leqq \eta \leqq 1$. These solutions form, for given m, a countable sequence of "angular wave functions of the first kind" $S_{mn}^{(1)}(c, \eta)$, $n = 0, 1, \cdots$. The solutions of (3) for the same values of λ_{mn} and which are also free of logarithmic terms are chosen as the "radial wave functions of the first kind" $R_{mn}^{(1)}(c, \xi)$. Since only functions of the first kind will be treated here, and since the parameter c will remain fixed in the discussion, the superscript (1) and the argument c will be omitted. The properties of these functions are easily deduced from explicit series expansions given in [1]. Page references to this report are given below.

The angular wave functions have the following behavior at the end points of their interval of definition: as $\eta \to 1$,

$$S_{mn}(\eta) = 0(1-\eta),\ dS_{mn}(\eta)/d\eta = 0(1),\ m \neq 0, 1;$$

$$S_{on}(\eta) = 0(1),\ dS_{on}(\eta)/d\eta = 0(1); \tag{4}$$

$$S_{1n}(\eta) = (1-\eta)^{\frac{1}{2}}[a_n + 0(1-\eta)], \qquad dS_{1n}(\eta)/d\eta \sim b_n(1-\eta)^{-\frac{1}{2}}$$

[p. 32]. The behavior as $\eta \to -1$ is analogous, with $1-\eta$ replaced by $1+\eta$.

The radial functions behave as follows at the finite end of their interval: as $\xi \to 1$,

$$R_{mn}(\xi) = 0(\xi - 1), \qquad dR_{mn}(\xi)/d\xi = 0(1), \qquad m \neq 0, 1;$$

$$R_{on}(\xi) = 0(1), \qquad dR_{on}(\xi)/d\xi = 0(1); \tag{5}$$

$$R_{1n}(\xi) = (\xi-1)^{\frac{1}{2}}[a'_n + 0(\xi-1)], \qquad dR_{1n}(\xi)/d\xi \sim (\xi-1)^{-\frac{1}{2}}[b'_n + 0(\xi-1)]$$

[pp. 34–35]. As $\xi \to \infty$, the asymptotic behavior is described by:

$$R_{mn}(\xi) \sim (c\xi)^{-1}\cos[c\xi - \tfrac{1}{2}(n+1)\pi],$$

$$dR_{mn}(\xi)/d\xi \sim -c(c\xi)^{-1}\sin[c\xi - \tfrac{1}{2}(n+1)\pi], \tag{6}$$

[p. 24.]

The existence of the integrals

$$\int_1^\infty R_{mn}(\xi)R_{mp}(\xi)\,d\xi$$

is easily established from (5) and (6). For $n \neq p$, the standard Sturm-Liouville method leads to the partial orthogonality relations

$$\int_1^\infty R_{mn}(\xi)R_{mp}(\xi)\,d\xi = \begin{cases} 0, & n-p \text{ even}, \\ (-1)^{(n-p-1)/2}/c(\lambda_{mn} - \lambda_{mp}), & n-p \text{ odd} \end{cases} \tag{7}$$

[cf. 3, p. 237–238]; the behavior at infinity detailed in (6) gives the stated values for the integrated term. The case $n = p$ gives the normalization integral

$$\int_1^\infty R_{mn}^2(\xi)\,d\xi = M_{mn}. \tag{8}$$

Similarly, one finds that

$$\int_{-1}^1 S_{mn}(\eta)S_{mp}(\eta)\,d\eta = \delta_n^p N_{mn}, \tag{9}$$

where δ_n^p is the Kronecker δ symbol.

3. Expansion of Solutions. The methods described in [2, Ch. IV] or [5, Ch. VI, §4], and in [3, Ch. XI], show that an arbitrary regular solution of the scalar wave equation may be expressed in a termwise integrable double series of the functions

$$\psi_{mn}(\eta, \xi, \phi) = S_{mn}(\eta)R_{mn}(\xi)\cos m\phi$$

and

$$\omega_{mn}(\eta, \xi, \phi) = S_{mn}(\eta)R_{mn}(\xi) \sin m\phi.$$

Write $\psi_{mn}(\eta, \xi, \phi) = \Psi_{mn}(x, y, z)$. The functions $\partial\Psi_{mn}/\partial x$, $\partial\Psi_{mn}/\partial y$, $\partial\Psi_{mn}/\partial z$ are continuous solutions of the scalar wave equation (1). The derivatives $\partial\Psi_{mn}/\partial x$ and $\partial\Psi_{mn}/\partial z$ are even functions of ϕ and may be expanded in a series of functions $\Psi_{\mu\nu}$, while $\partial\Psi_{mn}/\partial y$ is an odd function of ϕ and may be expanded in a series of functions $\omega_{\mu\nu}$. The first set of formulas will be obtained in detail from the expansion

$$\partial\Psi_{mn}/\partial z = \sum_{\mu=0}^{\infty}\sum_{\nu=\mu}^{\infty} g_{\mu\nu}^{mn}\psi_{\mu\nu}(\eta, \xi, \phi). \tag{10}$$

For the other two sets of formulas only an indication of certain modifications of the method and the final results will be given.

Expansion of the left member of (10) gives:

$$[\xi(1 - \eta^2)S'_{mn}R_{mn} \cos m\phi + \eta(\xi^2 - 1)S_{mn}R'_{mn} \cos m\phi]/d(\xi^2 - \eta^2) = \sum_{\mu=0}^{\infty}\sum_{\nu=\mu}^{\infty} g_{\mu\nu}^{mn}S_{\mu\nu}(\eta)R_{\mu\nu}(\xi) \cos \mu\phi. \tag{11}$$

Multiply both members of (11) by $\cos p\phi$ and integrate from 0 to 2π. On the left there is a non-zero result only for $p = m$, on the right only for $p = \mu$:

$$0 = \sum_{\nu=p}^{\infty} g_{p\nu}^{mn}S_{p\nu}(\eta)R_{p\nu}(\xi), \qquad p \neq m \tag{12}$$

and

$$[\xi(1 - \eta^2)S'_{mn}R_{mn} + \eta(\xi^2 - 1)S_{mn}R'_{mn}]/d(\xi^2 - \eta^2) = \sum_{\nu=m}^{\infty} g_{m\nu}^{mn}S_{m\nu}R_{m\nu}. \tag{13}$$

Next multiply by $S_{p\kappa}(\eta)$ or $S_{m\kappa}(\eta)$, respectively, integrate from -1 to 1 with respect to η, and apply (9). From (12) one obtains

$$g_{p\nu}^{mn} = 0, \qquad p \neq m,$$

i.e. in the expansion of $\partial\Psi_{mn}/\partial z$ only the functions $\psi_{m\nu}$ need have been used. In (13) there is a non-zero result on the right only for $\kappa = \nu$. The left member is the sum of the integrals

$$I_{\nu}^{mn}(\xi) = \frac{2}{d}\int_{-1}^{1} R_{mn}(\xi)\xi\,\frac{1 - \eta^2}{\xi^2 - \eta^2}\,S'_{mn}(\eta)S_{m\nu}(\eta)\,d\eta$$

$$J_{\nu}^{mn}(\xi) = \frac{2}{d}\int_{-1}^{1} R'_{mn}(\xi)\eta\,\frac{\xi^2 - 1}{\xi^2 - \eta^2}\,S_{mn}(\eta)S_{m\nu}(\eta)\,d\eta. \tag{14}$$

The existence of these integrals must be demonstrated. For $m \neq 1$, the only difficulty which might arise comes from the functions $(1 - \eta^2)/(\xi^2 - \eta^2)$ and $(\xi^2 - 1)/(\xi^2 - \eta^2)$ near $\eta = 1$, $\xi = 1$, or $\eta = -1$, $\xi = 1$. However, since the functions are bounded near these points, the integrals exist. For $m = 1$, by virtue of (4), (5) the integrand of I near $\eta = 1$, $\xi = 1$ may be written as:

$$b_n\xi(\xi - 1)^{\frac{1}{2}}[a'_n + 0(\xi - 1)]\,\frac{1 - \eta}{\xi - \eta}\,\frac{1 + \eta}{\xi + \eta}\,[a_\nu + 0(1 - \eta)]$$

while the integrand of J has the form:

$$(\xi - 1)^{\frac{1}{2}}[b'_n + 0(\xi - 1)]\eta \frac{\xi + 1}{\xi + \eta} \frac{1 - \eta}{\xi - \eta} [a_n a_r + 0(1 - \eta)].$$

Similar estimates hold near $\eta = -1, \xi = 1$. The integrands are again seen to be bounded, and the integrals converge.

The series representations in [1, p. 32] show that the integrands in the integrals I, J, are odd functions of η for $n - m$ and $\nu - m$ both even or both odd, and the coefficients g_{ν}^{mn} are accordingly zero when $\nu - n$ is an even integer or zero.

4. Recurrence Relations for Radial Functions. The formulas obtained can now be written as

$$(15) \quad (2/d) \int_{-1}^{1} (\xi^2 - \eta^2)^{-1}[\xi(1 - \eta^2)S'_{mn}(\eta)S_{m\nu}(\eta)R_{mn}(\xi) + \eta(\xi^2 - 1)S_{mn}(\eta)S_{m\nu}(\eta)R'_{mn}(\xi)]\, d\eta = g_{\nu}^{mn} N_{m\nu} R_{m\nu}(\xi),$$

for $n = 0, 1, \cdots$, and for $\nu - n$ odd. The relations in (15) display $R_{m\nu}$ as a linear combination of R_{mn} and R'_{mn}, with coefficients that are continuous functions of ξ in any finite subinterval of the range $1 \leqq \xi < \infty$.

For evaluation of the constant g_{ν}^{mn}, an integration over ξ is required. For large ξ, the significant factors in I_{ν}^{mn} are

$$(\xi^2 - \eta^2)^{-1}\xi R_{mn}(\xi) = 1/\xi(1 - \eta^2/\xi^2)^{-1}R_{mn}(\xi),$$

while those in J_{ν}^{mn} are

$$(\xi^2 - 1)(\xi^2 - \eta^2)^{-1}R'_{mn}(\xi) = (1 - \eta^2/\xi^2)^{-1}(1 - 1/\xi^2)R'_{mn}(\xi).$$

It follows from (6) that the integrals converge at infinity. Near $\xi = 1$, the estimates established previous to the integration with respect to η are adequate to prove the existence of the integrals. The integration yields

$$(16) \quad g_{\nu}^{mn} = 2/dN_{m\nu}M_{m\nu} \int_1^{\infty} \int_{-1}^{1} (\xi^2 - \eta^2)^{-1}[\xi(1 - \eta^2)R_{mn}R_{m\nu}S'_{mn}S_{m\nu} + \eta(\xi^2 - 1)R'_{mn}R_{m\nu}S_{mn}S_{m\nu}]\, d\eta\, d\xi.$$

Alternatively, after dividing both sides of (15) by $N_{m\nu}R_{m\nu}(\xi)$ one may note that the right hand member is constant. Therefore the left hand member is independent of ξ, and substitution of a convenient value ξ_0 gives g_{ν}^{mn} in terms of the integral over η.

5. Recurrence Relations for Index m. Relations connecting the radial functions $R_{mn}(\xi)$, $R'_{mn}(\xi)$, and $R_{m\pm1,\nu}(\xi)$ are obtained by an application of the method to the expansion of $\partial\Psi_{mn}/\partial x$ in functions $\psi_{\mu\nu}(\eta, \xi, \phi)$. The expansion of $\partial\Psi_{mn}/\partial y$ in functions $\omega_{\mu\nu}(\eta, \xi, \phi)$ would lead to the same relations. Similarly to (11), the expansions give

$$\partial\Psi_{mn}/\partial x = \sum_{p=0}^{\infty}\sum_{\nu=0}^{\infty} h_{p\nu}^{mn}\psi_{p\nu}$$

$$= (2T/dD)[-\eta S'_{mn}R_{mn} + \xi S_{mn}R'_{mn}]\cos\phi\cos m\phi \tag{17}$$

$$+ (2m/dT)S_{mn}R_{mn}\sin\phi\sin m\phi,$$

where

$$D = \xi^2 - \eta^2,\ T = (1-\eta^2)^{\frac{1}{2}}(\xi^2-1)^{\frac{1}{2}}. \tag{18}$$

Trigonometric transformation changes the right hand members to linear combinations of $\cos(m \pm 1)\phi$, and multiplication by $\cos p\phi$ and integration with respect to ϕ from 0 to 2π eliminates one summation:

$$\sum_{\nu=0}^{\infty} h_{p\nu}^{mn}S_{p\nu}R_{p\nu} = 0, \qquad m > 0, \qquad p \neq m \pm 1;$$
$$\sum_{\nu=0}^{\infty} h_{p\nu}^{on}S_{p\nu}R_{p\nu} = 0, \qquad p \neq 1, \tag{19}$$

and

$$\sum_{\nu=0}^{\infty} h_{m\pm1,\nu}^{mn}S_{m\pm1,\nu}R_{m\pm1,\nu} = \mp\frac{T}{dD}(\eta S'_{mn}R_{mn} - \xi S_{mn}R'_{mn})$$

$$-\frac{m}{dT}S_{mn}R_{mn}, \qquad m > 0, \tag{20}$$

$$\sum_{\nu=0}^{\infty} h_{1\nu}^{on}S_{1\nu}R_{1\nu} = -\frac{2T}{dD}(\eta S'_{on}R_{on} - \xi S_{on}R'_{on}).$$

The procedure of section 3 gives the following recurrence relations for the radial functions:

$$h_{m\pm1,\nu}^{mn}N_{m\pm1,\nu}R_{m\pm1,\nu} = \mp(1/d)\int_{-1}^{1}(T/D)(\eta S'_{mn}R_{mn} - \xi S_{mn}R'_{mn})S_{m\pm1,\nu}\,d\eta.$$

$$-(m/d)\int_{-1}^{1}(1/T)S_{mn}R_{mn}\,d\eta, \qquad m > 0, \tag{21}$$

$$h_{1\nu}^{on}N_{1\nu}R_{1\nu} = -(2/d)\int_{-1}^{1}(T/D)(\eta S'_{on}R_{on} - \xi S_{on}R'_{on})S_{m\pm1,\nu}\,d\eta.$$

These relations are valid for $n = 0, 1, \cdots$, and for $n - \nu$ an odd integer. The coefficients corresponding to $\nu - n$ zero or even, as well as all coefficients not appearing in formulas (21), are zero. The evaluation of the non-zero coefficients $h_{p\nu}^{mn}$ may be carried out as suggested in section 4.

6. Recurrence Relations for Separation Constants. Because of the incomplete orthogonality (7) of the radial functions, the procedure of the foregoing sections does not appear to yield recurrence relations for the angular functions $S_{mn}(\eta)$. However, a related procedure can be used to derive a recurrence relation for the (non-integral) separation constants λ_{mn}. Estimates similar to those in the discussions of the integrals (14) and (16) show that multiplication of both members

of (13) by $R_{m\nu}(\xi)$ and integration with respect to ξ from 1 to ∞ is a valid operation. The result is the formula

$$(22)\qquad (2/d)\int_1^\infty (1/D)[\xi(1-\eta^2)S'_{mn}R_{m\nu}+\eta(\xi^2-1)S_{mn}R'_{mn}R_{m\nu}]\,d\xi$$
$$= g^{mn}_{m\nu}M_{m\nu}S_{m\nu}+\sum{}'^{\infty}_{p=o,l}(-1)^{(\nu-p-1)/2}g^{mn}_{mp}S_{mp}/c(\lambda_{mp}-\lambda_{m\nu}),$$

where $D=\xi^2-\eta^2$ as before, and where the summation symbol denotes a sum over even p if ν is odd, a sum over odd p if ν is even. Next both members are multiplied by $S_{mp}(\eta)$ and integrated with respect to η from -1 to 1 to produce the relation

$$d^{mn}_{p\nu}=(-1)^{(\nu-p-1)/2}g^{mn}_{mp}N_{mp}/c(\lambda_{mp}-\lambda_{m\nu}),$$

where $d^{mn}_{p\nu}$ stands for the integral of the left member of (22) with respect to η. The ratio of $d^{mn}_{p\nu}$ and g^{mn}_{mp} is evidently independent of n. Writing

$$c^m_{p\nu}=(-1)^{(\nu-p-1)/2}g^{mn}_{mp}N_{mp}/cd^{mn}_{p\nu},$$

one has finally

$$(23)\qquad \lambda_{mp}-\lambda_{m\nu}=c^m_{p\nu}$$

for any integers p, ν whose difference is an odd integer. More generally,

$$(24)\qquad \lambda_{mn}=\lambda_{mo}+\sum{}^n_{p=l}c^m_{p,p-l}\,.$$

A similar procedure applied to the formulas of section 5 leads to no essentially different results.

Some further details about the coefficients in the recurrence relations will appear in UMM 126, a forthcoming publication of the Willow Run Research Center.

REFERENCES

1. C. Flammer, "Spheroidal Wave Functions," Technical Report, Stanford Research Institute, Stanford, 1951.
2. B. Friedman, "Techniques in Solving Partial Differential Equations," Institute for Mathematics and Mechanics, New York University, 1950.
3. E. L. Ince, "Ordinary Differential Equations," Dover, New York, 1944.
4. E. Madelung, "Die Mathematischen Hilfsmittel des Physikers," Dover, New York, 1943.
5. P. M. Morse and H. Feshbach, "Methods of Theoretical Physics," Technology Press, M. I. T., 1946.
6. J. A. Stratton, P. M. Morse, L. J. Chu, and R. A. Hutner, "Elliptic Cylinder and Spheroidal Wave Functions," Wiley, New York, 1950.
7. E. T. Whittaker, "Recurrence Formulae for Mathieu Functions," Jour. L. M. S. 4, Vol. 88 (1929).

Willow Run Research Center of the University of Michigan

(Received January 12, 1953; Revised June 15, 1953)

DYNAMIC ANALYSIS OF THE BUCKLING OF LATERALLY LOADED FLAT ARCHES*

By N. J. Hoff and Victor G. Bruce

Introduction.[1] There is a type of buckling which sometimes occurs when curved elastic elements are so loaded that the load tends to reduce the curvature. The buckling consists of a sudden change of curvature or "jump" of the element and, upon subsequent removal of the load, the element may, or may not, return to its original configuration. A familiar example, in which the element is a shell, is the convex bottom of an oil can. Another example, which is simpler from the standpoint of analysis, and which illustrates the essential features of the phenomenon, is the laterally loaded, slightly curved bar or arch. It is this example that will be treated here.

This type of buckling, called "Durchschlag" in German literature, has been discussed by a number of authors. The significance of the problem, in so far as it illustrates certain important features in more complicated buckling problems of plates and shells, was pointed out by Marguerre (1, 2)[2], who constructed a simplified model to demonstrate these features. The question was also discussed by Timoshenko (3, 4) and by Biezeno (5, 6). More recently, the problem has been treated by the authors of this paper (7, 8), and by Y. C. Fung (9).

In the present paper the problem is approached from the dynamical point of view, in contrast to the usual statical treatment. This permits a more rigorous analysis of the stability under constant load. In addition, one can calculate certain other dynamic features of interest, such as the critical value of a suddenly applied load of infinite duration. Although both the shape of the bar and the lateral load distribution are assumed to be sinusoidal, the results are qualitatively and, within limits, quantitatively applicable to cases in which both the shape and load distribution deviate from these assumptions. This can be shown by comparing the results of this analysis with those obtained by Biezeno, who considered a bar of parabolic shape loaded at the mid-point with a concentrated load.

Problem and assumptions. A slightly curved bar, free to move in the horizontal (x, y) plane, is acted on by a distributed time dependent load $Q(x, t)$ as shown in Figure 1. The bar is pinned at its ends, which are held a fixed distance L apart. The curve $y = y_0(x)$ defines the unloaded equilibrium position, and in this position the support reactions vanish. It is desired to find the motion of the bar.

It is assumed that the curvature is small and that the axial stress is constant

* Paper presented before the Eighth International Congress on Theoretical and Applied Mechanics in Istanbul, Turkey, August 20–28, 1952.

[1] The contents of this paper are based on research carried out at the Polytechnic Institute of Brooklyn under Contract No. Nonr-267(00), Project No. NR 064-363 sponsored by the Office of Naval Research of the U. S. Navy.

[2] Numbers in parentheses refer to the bibliography.

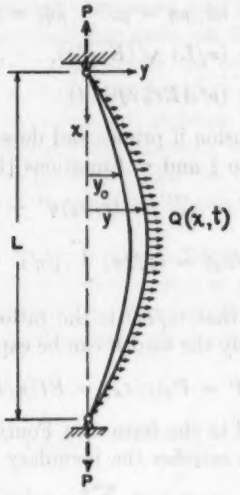


FIG. 1 ARCH DEFLECTED UNDER LOAD

along the bar and is thus a function only of the time t. The effects of rotatory inertia and shearing deformations are neglected and damping forces are excluded. Strains are assumed to remain within the elastic limit and Hooke's law is valid. The bar is uniform in cross section with cross section A, volume mass density σ, and with one principal axis lying in the plane of motion. I is the moment of inertia of the section.

Derivation of the equations. Because of the assumed small curvature, and constancy of the axial load P with respect to the x coordinate, the differential equation of motion can be written as

$$EI(y'''' - y_0'''') - Py'' + \sigma A\ddot{y} - Q(x, t) = 0 \qquad [1]$$

where the axial load P is

$$P = (1/2)(EA/L) \int_0^L [(y')^2 - (y_0')^2]\, dx \qquad [2]$$

The primes indicate differentiation with respect to x, and the dots with respect to the time t. These equations may be put into dimensionless form by means of the substitutions

$$L\xi = nx,\ \rho\eta = y, \qquad \rho\eta_0 = y_0\text{, where } A\rho^2 = I;$$
$$(\tau/t) = (\pi/L)\sqrt{(E\varepsilon_E/\sigma)}, \qquad \text{where } \varepsilon_E = (\pi\rho/L)^2; \tag{3}$$
$$q(\xi, \tau) = (\rho/AE\varepsilon_E^2)Q(x, t).$$

It should cause no confusion if primes and dots are again used to denote differentiation with respect to ξ and τ. Equations [1] and [2] then become

$$\eta'''' - \eta_0'''' - (\varepsilon/\varepsilon_E)\eta'' + \ddot{\eta} - q = 0, \tag{4}$$

$$(\varepsilon/\varepsilon_E) = (1/2\pi)\int_0^\pi [(\eta')^2 - (\eta_0')^2]\, d\xi \tag{5}$$

It should be remarked that $(\varepsilon/\varepsilon_E)$ is the ratio of the axial strain to the first Euler strain, consequently the load P can be expressed by means of the relation,

$$P = P_E(\varepsilon/\varepsilon_E) = EI(\pi/L)^2(\varepsilon/\varepsilon_E)$$

A solution is assumed in the form of a Fourier series with coefficients which depend on τ and which satisfies the boundary conditions:

$$\eta = \eta_0 + \sum_{n=1}^{\infty} a_n(\tau) \sin n\xi. \tag{6}$$

It is also assumed that

$$\eta_0 = e \sin \xi, \text{ so that } \rho e = (y_0)_{\max},$$

and

$$q = \sum_{n=1}^{\infty} q_n(\tau) \sin n\xi$$

With these assumptions the variables can be separated. Because of the orthogonality relations for the sines, Equation [5] simplifies to

$$(\varepsilon/\varepsilon_E) = (1/4)\sum_{n=1}^{\infty} n^2 a_n^2 + 2ea_1.$$

Equation [4] becomes,

$$\sum_{n=1}^{\infty} n^4 a_n \sin n\xi + (\varepsilon/\varepsilon_E)\Big(\sum_{n=1}^{\infty} n^2 a_n \sin n\xi + e \sin \xi\Big) + \sum_{n=1}^{\infty} \ddot{a}_n \sin n\xi = \sum_{n=1}^{\infty} q_n \sin n\xi.$$

By successively equating to zero the coefficients of $\sin \xi$, $\sin 2\xi$, $\cdots$, etc., one obtains the following set of equations:

$$\ddot{a}_1 + a_1 + (1/4)\Big(\sum_1^{\infty} n^2 a_n^2 + 2ea_1\Big)(a_1 + e) = q_1 \tag{7.1}$$

$$\ddot{a}_2 + 2^4 a_2 + (1/4)\Big(\sum_1^{\infty} n^2 a_n^2 + 2ea_1\Big)2^2 a_2 = q_2 \tag{7.2}$$

$$\cdots$$

$$\ddot{a}_k + k^4 a_k + (1/4)\Big(\sum_1^{\infty} n^2 a_n^2 + 2ea_1\Big)k^2 a_k = q_k \tag{7.3}$$

$$\cdots$$

Equilibrium under quasi-static half-sine-wave load distribution. If the lateral load is distributed according to a half sine wave, all the q_k except q_1 are identically zero. If, furthermore, $q_1(\tau)$ varies slowly (as compared to the period of small oscillations of the bar), such loading is called quasi-static. Under these conditions the problem is a static equilibrium problem. If, then the differentiated terms in Equations [7 $\cdots$] are omitted, solutions of the resulting algebraic equations represent equilibrium positions under the load q_1.

Evidently there are two possible solutions of this set of equations. In the first solution, $a_1 \neq 0$ while $a_2 = a_3 = \cdots = a_k = \cdots = 0$. In the second solution, $a_1 \neq 0$, $a_k \neq 0$ while $a_n = 0$ if $n \neq 1$ or k. Furthermore, it is not possible to have a solution in which $a_k \neq 0$, $a_j \neq 0$, $k \neq j \neq 1$, for then the two equations in k and j would be contradictory. There are thus two possible solutions determining the equilibrium configurations corresponding to different values of q_1 and e.

In the first case,

$$a_1 + (\tfrac{1}{4})(a_1^2 + 2ea_1)(a_1 + e) = q_1\,, \qquad a_k = 0 \tag{8}$$

In the second case,

$$\begin{aligned} a_1 + (\tfrac{1}{4})(a_1^2 + k^2a_k^2 + 2ea_1)(a_1 + e) &= q_1\,, \\ k^2 + (\tfrac{1}{4})(a_1^2 + k^2a_k^2 + 2ea_1) &= 0. \end{aligned} \tag{9}$$

A more symmetrical result is obtained through the introduction of the notation $a_1 = e(z - 1)$, $a_k = ex_k$, and $q_1 = e(w - 1)$. Then the equations become

$$(e/2)^2z[z^2 - 1 + (2/e)^2] = w, \qquad x_k = 0; \tag{10}$$

and

$$\begin{aligned} (e/2)^2z[z^2 - 1 + (2/e)^2] &= w - k^2(e/2)^2x_k^2z\,, \\ e^2[z^2 - 1 + k^2(2/e)^2] + e^2k^2x_k^2 &= 0. \end{aligned}$$

The physical significance of the new variables can be directly found from the definitions of the non-dimensional quantities. One has,

$$z = y_{1\cdot\max}/y_{0\cdot\max}.$$

This is the ratio of the amplitude of the first harmonic of the total displacement to the amplitude of the initial displacement. Similarly

$$x_k = y_{k\cdot\max}/y_{0\cdot\max}$$

is the ratio of the amplitude of the kth harmonic of the total displacement to the amplitude of the initial displacement. Finally, the non-dimensional load parameter (half sine wave distribution) can be expressed as

$$w = (\tfrac{1}{2}\pi)(L/y_{0\cdot\max})(P_{tr}/P_E) + 1,$$

where P_{tr} is the total transverse load,

$$P_{tr} = (2/\pi)LQ_{\max}.$$

The unloaded condition, $Q = 0$, corresponds to $w = 1$. When w vanishes, the total transverse load is

$$P_{tr} = -2\pi(y_{0\cdot max}/L)P_E.$$

The two equations following Equations [10] may be written in an equivalent form, in which it is evident that the result holds only when $e \geqq 4$.

$$(1 - k^2)z = w, \qquad z^2 + k^2x_k^2 = 1 - k^2(2/e)^2, \qquad k = 2, 3 \cdots . \qquad [11]$$

Equation [10] has been solved for values of $e = 1, 2, 3$ and 4, and is plotted in Figure 2. Evidently, all curves pass through the points $(z, w) = (0, 0), (+1, +1)$

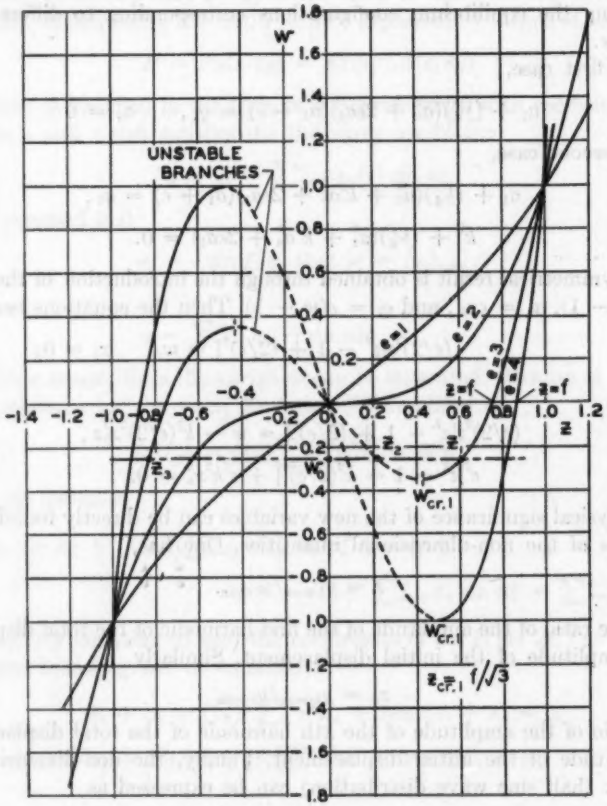


FIG. 2 EQUILIBRIUM POSITIONS UNDER QUASI-STATIC LOAD
$e \leq 4$

and $(-1, -1)$. The results represent load-deflection characteristics of the arch for various values of the parameter e, which is a measure of the initial eccentricity of the arch.

Not all of these equilibrium positions are stable, however, and the question of stability will now be discussed in a little more detail. If the acceleration terms are retained, the considerations and substitutions that led earlier to Equations [9] result in the following simultaneous equations:

$$\begin{aligned}(2/e)^2\ddot{z} + [z^2 - 1 + (2/e)^2]z - (2/e)^2 w + k^2 x_k^2 z = 0,\\ \ddot{x}_k + k^2[k^2 + (e/2)^2(z^2 - 1 + k^2 x_k^2)]x_k = 0.\end{aligned} \tag{12}$$

Suppose now that, $z = z_1$, z_2 or z_3 and $x_k = 0$ are the equilibrium positions corresponding to w_0 as given by Equation [10], (Figure 2). To examine the stability of any such equilibrium position, disturbances (or variations) $\mu(\tau)$ and $\nu(\tau)$ are introduced.

$$\begin{aligned}z &= z_i + \mu(\tau), \qquad i = 1, 2 \text{ or } 3.\\ x_k &= \nu(\tau).\end{aligned} \tag{13}$$

The values of w_0 and z_i are constant, while μ and ν are small deviations from the equilibrium position. Introducing [13] into [12], and retaining only first order terms, one gets

$$\begin{aligned}(2/e)^2\ddot{\mu} + [3z_i^2 - 1 + (2/e)^2]\mu = 0\\ (2/e)^2\ddot{\nu} + k^2[z_i^2 + k^2(2/e)^2 - 1]\nu = 0 \qquad (k \geqq 2)\end{aligned} \tag{14}$$

If $e \leqq 2$, the coefficients of μ and ν in [14] are never negative and the equilibrium is always stable. If $e > 2$, and if the abbreviation $f^2 = 1 - (2/e)^2$ is used, it is seen that the equilibrium is unstable when $3z_i^2 < f^2$. It is easy to show that this inequality holds for the root z_2, (see Figure 2). In fact, one can prove that the value of z at which the curves have horizontal tangents is just $z = \pm z_{cr\cdot 1} = \pm f/\sqrt{3}$. The corresponding value of w is $w_{cr\cdot 1} = \pm(2f^3/3\sqrt{3})(e/2)^2$. The equilibrium represented by Equations [14] is also unstable when $z_i^2 < 1 - k^2(2/e)^2$. Since $k \geqq 2$, this inequality can only hold when $e > 4$.

From the preceeding it may be seen that if $e \leqq 2$, no sudden jump can occur. If $2 < e \leqq 4$, a sudden jump takes place when w reaches the value $w_{cr\cdot 1}$, and the only stable equilibrium configurations under the load $w = w_0$ are those corresponding to $z = z_1$ or z_3, (see Figure 2). If $e > 4$ one must consider also the solution given in Equation [11].

In Figure 3, for example, there has been plotted the equilibrium configurations for the case when $e = 8$. The solution corresponding to Equation [10] is shown as well as the solution corresponding to Equation [11], with $k = 2$. To examine the stability of the second solution one must use the following, instead of [13], in the dynamical equations [12]:

$$z = -w_0/3 + \mu(\tau), \qquad x = x_i + \nu(\tau), \qquad i = 1 \text{ or } 2. \tag{15}$$

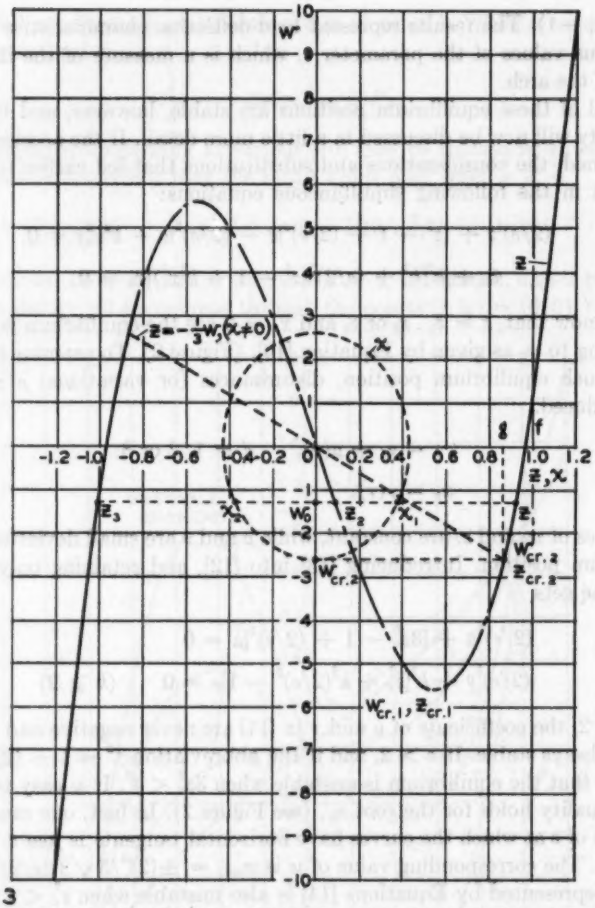


FIG. 3
EQUILIBRIUM POSITIONS UNDER QUASI-STATIC LOAD
$e = 8$

The equilibrium values for z and x, i.e. $-w_0/3$ and x_i, are those obtained from Equation [11], with $k = 2$. The notation $g^2 = 1 - (4/e)^2$ is introduced; so that the intersection point, $(z_{cr\cdot 2}, w_{cr\cdot 2})$, has the coordinates $(g, -3g)$.

The equations for the disturbances, μ and ν, are, to the first order in μ and ν,

$$\ddot{\mu} + [(2/9)(e/2)^2 w_0^2 - 3]\mu = (2/3)e^2 w_0 x_i \nu,$$
$$\ddot{\nu} + 8e^2 x_i^2 \nu = (2/3)e^2 w_0 x_i \mu. \qquad [16]$$

These can be written in the following form:

$$\ddot{\mu} + a\mu = b\nu, \qquad \ddot{\nu} + c\nu = b\mu$$

The characteristic equation, obtained by letting $\mu = M \exp \lambda t$, $\nu = N \exp \lambda t$, and solving the resulting homogeneous linear equations, is

$$\lambda^4 + (a + c)\lambda^2 + (ac - b^2) = 0.$$

Since $ac - b^2 = -24e^2x_1^2 \leqq 0$, the equilibrium represented by these equations is always unstable, i.e. the second mode is only temporary.

By equating $z_{cr\cdot 1}$ to $z_{cr\cdot 2}$, i.e. $f/\sqrt{3} = g$, one obtains $e = \sqrt{22}$. Consequently, all equilibrium configurations characterized by $z = z_1$, $x = 0$ are stable if $4 < e < \sqrt{22}$ (see Equation 14). If $e > \sqrt{22}$ some equilibrium configurations $z = z_1$, $x = 0$ are not stable.

Figure 3 includes the result of these stability calculations for the case $e = 8$. Unstable equilibrium configurations are shown in dashed lines and stable configurations in full lines. It may be noted that if $4 < e < \sqrt{22}$ the straight line $z = -w/3$ intersects the cubic curve to the left of the point $z_{cr\cdot 1}$, whereas if $e > \sqrt{22}$ the intersection lies to the right, as shown in Figure 3.

If the jump is associated with instability of the equilibrium $z = z_1$, $x = 0$, it follows that if $e < 2$, no jump occurs. If $2 < e < \sqrt{22}$, the critical load is $w_{cr\cdot 1}$. If $e > \sqrt{22}$ the critical load is $w_{cr\cdot 2}$. Furthermore, it is seen that if $2 < e < 4$ a jump in the first mode alone occurs, whereas if $e > 4$ deflections in both the first and second modes take place.

The transient problem with a half-sine-wave distributed load. Although the quasi-static treatment is sufficient to enable one to calculate buckling loads when the load is applied slowly, there are other problems of interest. For instance, will a suddenly applied sub-critical load cause buckling or will a short-duration super-critical load lead to failure? An explicit analytical solution of this problem does not seem to be possible, but some results of interest can be obtained for a special type of load application if some simplifying assumptions are introduced.

The load, $q(\xi, \tau)$, has been assumed as a step function of magnitude q_1 and duration $\Delta\tau (q_1 < 0)$:

$$q(\xi, \tau) = \begin{cases} q_1 \sin \xi, & 0 < \tau < \Delta\tau, \\ 0 & \tau < 0 \text{ and } \tau > \Delta\tau. \end{cases}$$

One may obtain an energy integral by assuming that only the first and second modes can be different from zero. Thus, in Equations [12] let $x_k = x_2$, in which case

$$(2/e)^2\ddot{z} + (z^2 - 1 + 4/e^2)z - (2/e)^2 w + 4x_2^2 z = 0$$
$$\ddot{x}_2 + 4[4 + (e/2)^2(z^2 - 1 + 4x_2^2)]x_2 = 0$$

If the first equation is multiplied by $\dot{z}(e/2)^2\, d\tau$, the second by $\dot{x}_2\, d\tau$, and if the equations are added and integrated the result is

$$\dot{z}^2 + \dot{x}_2^2 + (\tfrac{1}{4})(4 - e^2)z^2 + (16 - e^2)x_2^2 + (e^2/8)z^4$$
$$+ 2e^2x_2^4 + e^2x_2^2z^2 = 2\int w\, dz + E'$$

where E' is a constant of integration which depends on the initial conditions. If $r_1 = ez$, $a_2 = ex_2$, and $Q = ew$ = constant, then the above equation becomes

$$\dot{r}_1^2 + \dot{a}_2^2 + (1/8)(r_1^2 + 4 - e^2)^2 - 2Qr_1 + a_2^2(2a_2^2 + r_1^2 + 16 - e^2) = E, \quad [17]$$

where E is a new integration constant.

Equation [17] is the energy integral of the system under a constant external load $Q - e = q_1$. The system is conservative and is dynamically essentially equivalent to a system consisting of a particle (of mass 2) sliding without friction on a surface $E(r_1, a_2)$ described by the equation

$$(1/8)(r_1^2 + 4 - e^2) - 2Qr_1 + a_2^2(2a_2^2 + r_1^2 + 16 - e^2) = E(r_1, a_2). \quad [18]$$

$E(r_1, a_2)$ represents the potential energy while $\dot{r}_1^2 + \dot{a}_2^2$ represents the kinetic energy.

In general the surface described by Equation [18] will contain "hills," "depressions," and "valleys" and the character of the surface will differ greatly if Q takes on different values. For a given value of Q the equilibrium positions will be located at points for which $\partial E/\partial r_1$ and $\partial E/\partial a_2$ both vanish. These points are the tops of the hills, the bottoms of the depressions, and the saddle points. In general, for a fixed value of Q there may be as many as five equilibrium positions. For example, if reference is made to Figure 3 it is seen that if $|w| < |w_{cr\cdot 2}|$ there are equilibrium positions at the following points (x, z): $(0, z_1)$ and $(0, z_3)$ are "depressions," $(0, z_2)$ is a "hill," and $(\pm x_1, -w/3)$ are "saddle points." The sketch in Figure 4 shows qualitatively the nature of the contours of constant potential energy of the surface $E(r_1, a_2)$ for this case.

When any initial conditions $r_1(0)$, $a_2(0)$, $\dot{r}_1(0)$, $\dot{a}_2(0)$ are prescribed, the motion of the point mass on the surface $E(r_1, a_2)$ is completely determined. However, small disturbances during the motion and slight deviations from the assumed initial conditions may occur with the result that the actual motion will be one in which the path followed will approach that one for which the slope of descent is everywhere a maximum. This "path of steepest descent" is, of course, well defined analytically but it has not been computed here. One can see from Figure 4 though that this path goes through the saddle point. Furthermore, of all paths connecting the two stable equilibrium positions it is the one along which the maximum potential energy has the smallest value. These considerations form the basis for the determination of the critical value of a suddenly applied, infinite duration, load Q_{cr}.

Suddenly applied load of infinite duration. A suddenly applied load of the type specified in the last section might be expected to cause buckling even though $|w| < |w_{cr\cdot 2}|$ provided $\Delta\tau$ is very large (infinite). Before applying the load the initial conditions are $\dot{a}_2(0) = \dot{r}_1(0) = a_2(0) = 0$, and $r_1(0) = e$. After the load is applied the energy surfaces will have the general character shown in Figure 4. The problem then is to find that value of $Q = Q_{cr}$ for which the potential contour passing through the point specified by the above initial conditions will also pass through the saddle point. If Q has this value one can say that

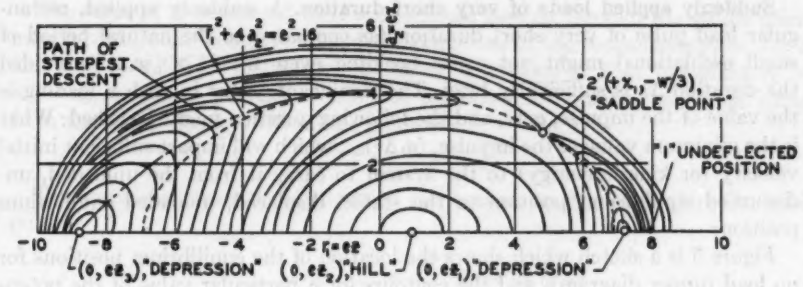


FIG. 4 SCHEMATIC SKETCH INDICATING CONTOURS OF EQUAL POTENTIAL ENERGY ($e = 8$). CURVES SHOWN ARE FOR THE CRITICAL VALUE OF A SUDDENLY APPLIED, INFINITE DURATION, LOAD.

it is possible for the initially undisturbed system to be displaced to the stable "buckled" equilibrium configuration, $(0, ez_3)$ if it follows the appropriate path on the energy surface.

Equation [17], with the initial conditions corresponding to the undisturbed state, gives

$$E = 2 - 2eQ_{cr}.$$

The contour of constant potential energy defined by Equation [18] when E has the above value must also pass through the saddle point, for which the coordinates are, $r_1 = -Q_{cr}/3$ and $a_2 = (1/2)\sqrt{e^2 - 16 - Q_{cr}^2/9}$. These values are, of course, obtained from Equations [11], where $k = 2$, if it is remembered that $r_1 = ez$, $a_2 = ex_2$, and $Q = ew$. The contour referred to is the one in Figure 4 which passes through point "1" (initial position) and point "2" (saddle point). Equation [18] thus becomes

$$(1/3)Q_{cr}^2 + 2eQ_{cr} - 32 + 3e^2 = 0.$$

The solution of this quadratic and selection of the appropriate root gives the result

$$q_{cr} = Q_{cr} - e = -4e(1 - \sqrt{6/e^2}) \qquad [19]$$

If $e = 8$ the critical value of a suddenly applied, infinite duration load is

$$q_{cr\,(\text{suddenly applied})} = -22.20.$$

This may be compared to the critical value of a slowly applied load (as given by $w_{cr\cdot 2}$) for which the value is

$$q_{cr\,(\text{slowly applied})} = -28.78$$

Suddenly applied loads of very short duration. A suddenly applied, rectangular load pulse of very short duration (as compared to the natural period of small oscillations) might not cause buckling even if $|w| > |w_{cr\cdot 2}|$ provided the duration $\Delta\tau$ is sufficiently brief. The significant factor in such a loading is the value of the impulse, $q_1\Delta\tau$, and the following question might be asked: What is the minimum value of the impulse, $(q_1\Delta\tau)_{cr}$, which will impart sufficient initial velocity (or kinetic energy) to the system to carry it from the unloaded, undisturbed equilibrium position to the stable, displaced, unloaded equilibrium position?

Figure 5 is a sketch which shows the location of the equilibrium positions for no load (upper diagram), and the contours for a particular value of the potential energy (lower diagram). The potential energy at ez_1 and ez_3 will be lower, whereas that at ez_2 will be higher than this value. The initial position is at ez_1.

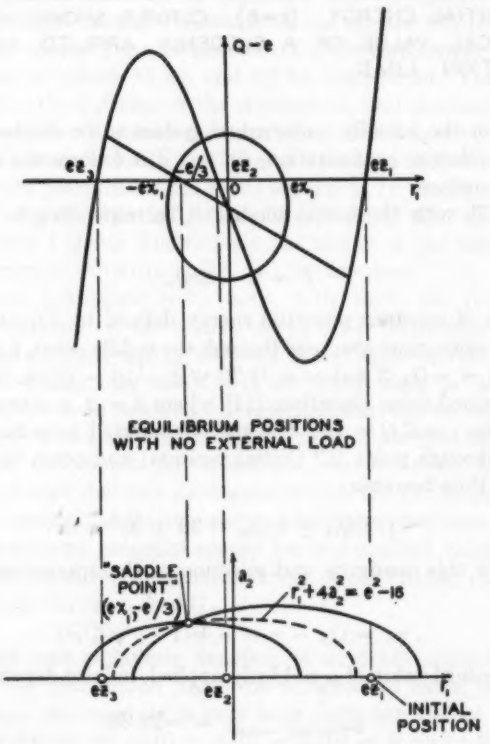


FIG. 5 CONTOURS OF EQUAL POTENTIAL ENERGY PASSING THROUGH THE SADDLE POINT

The value of E associated with the contours passing through the saddle point may be computed from [18] through introduction of the coordinates $r_1 = -e/3$ and $a_2 = \sqrt{2e^2/9 - 4}$ of the saddle point. The value of E at the point $r_1 = ez_1 = e$, $a_2 = 0$, is just $2(1 - e^2)$. The difference in potential energy in going from the initial point to the saddle point is

$$E = 16(e^2/3 - 2).$$

The impulsive velocity imparted to the system by the short-duration load pulse is

$$\dot{r}_{i_{(\text{imp})}} = e\dot{z}_{(\text{imp})} = \int_0^{\Delta\tau} e\ddot{z}\, d\tau \cong q_1 \Delta\tau$$

The kinetic energy associated with this velocity is $(q_1\Delta\tau)^2$.

If the kinetic energy acquired as a result of the impulse is equated to the increase in potential energy in going from ez_1 to the saddle point, a critical value of the impulse $(q_1\Delta\tau)_{cr}$ is defined which amounts to

$$(q_1\Delta\tau)_{cr} = -4\sqrt{e^2/3 - 2}$$

With $e = 8$

$$(q_1\Delta\tau)_{cr} = -17.52$$

To illustrate the significance of this result consider a value of $\Delta\tau = \tau_0/10$ where τ_0 is the natural period of small oscillations which, for $e = 8$, has the value $\tau_0 = 1.094$. The critical value of the load for this duration is thus -160 which is seen to be of much greater magnitude than the critical value of a slowly applied load.

Discussion and summary. It was shown in the paper that with very flat arches ($e \leqq 2$) no jump can occur. If the parameter e characterizing the arch rise ($e = y_0/\rho$) has a value between 2 and 4 a jump (buckling) can occur but there is only one unloaded equilibrium position. The critical load is $w_{cr\cdot1}$ (Figure 2). If $4 < e < \sqrt{22}$ the critical load is still given by $w_{cr\cdot1}$ but there are two stable unloaded equilibrium positions. If $e > \sqrt{22}$ the critical load is given by $w_{cr\cdot2}$ (Figure 3) and there are two stable unloaded equilibrium positions.

If $e > 4$ the problem can not be idealized as a system with only one degree of freedom and the shape during buckling may be nonsymmetrical even under symmetrical loading. One should note, however, that the final equilibrium buckled shape, either under load or with the load removed, is always symmetrical. This is an interesting result inasmuch as it shows that buckling loads which are calculated by assuming the deflected shape may be incorrect, even though the assumed deflected shape does accurately represent the final buckled configuration.

The critical value of a suddenly applied load of infinite duration was also calculated. For the particular case considered ($e = 8$) the value was found to

be about 25% less than the critical value of a slowly applied load. Loads of an impulsive type (suddenly applied, short duration) may, on the other hand, have values which are much greater than the critical value of a slowly applied load without precipitating a jump. The critical value of the impulse necessary to cause buckling was calculated.

Acknowledgment. The writers wish to acknowledge their indebtedness to the Office of Naval Research of the U. S. Navy for permission to present and publish this paper written as part of the research work carried out under the sponsorship of the Office. They are also indebted to James Sheng and William Berks for their assistance in checking the calculations as well as to Matthew G. Forte and Thomas J. Toomey for drawing the diagrams.

REFERENCES

(1) MARGUERRE, K., *Neuere Festigkeitsprobleme des Ingenieurs*, (edited by Marguerre), Springer, (1950), Ch. V.

(2) MARGUERRE, K., *Die Durchschlagskraft eines schwach gekrümmten Balkens*, Sitz. Berlin Math. Gess., **37**, (1938), p. 22.

(3) TIMOSHENKO, S. P., *Buckling of Curved Bars with Small Curvature*, J. Appl. Mech., **2**, (1935), p. 17.

(4) TIMOSHENKO, S. P., *Theory of Elastic Stability*, McGraw Hill, (1936), p. 230.

(5) BIEZENO, C. B., *Das Durchschlagen eines schwach gekrümmten Stabes*, Z. a. M. M., **18**, (1938), p. 21.

(6) BIEZENO-GRAMMEL, *Technische Dynamik*, Springer (1939), (Lithoprinted by Edwards Bros., 1944), p. 526.

(7) HOFF, N. J. AND BRUCE, VICTOR G., *Dynamic Analysis of the Buckling of Laterally Loaded Flat Arches*, Polytechnic Institute of Brooklyn Aeronautical Laboratories (PIBAL) Report No. 191, October, 1951.

(8) BRUCE, VICTOR G., *Further Investigations on the Problem of the Buckling of a Laterally Loaded Flat Arch*, Polytechnic Institute of Brooklyn Aeronautical Laboratories PIBAL Report No. 199, May 1952.

After completion of this paper the manuscript of a report dealing with the same subject but using the static equations of equilibrium was received. It will be published as an NACA Technical Note under the following title:

(9) FUNG, YUAN-CHENG, *Buckling of Low Arches or Curved Beams of Small Curvature.*

POLYTECHNIC INSTITUTE OF BROOKLYN

(Received September 26, 1952)

ON THE NUMERICAL SOLUTION OF VOLTERRA INTEGRAL EQUATIONS

By Carl Wagner

1. Introduction. Prasad[1], Carson[2], Huber[3], Collatz[4], and others have shown that Volterra integral equations may be solved numerically by a stepwise calculation. The integral equation may have the general form

$$\phi(x) = f(x) - g(x) \int_0^x \phi(z)K(x, z)\, dz \tag{1}$$

where $\phi(x)$ is an unknown function and $f(x)$, $g(x)$, and $K(x, z)$ are given functions. Using intervals of width h and letting $x = nh$ and $z = \nu h$, one may approximate the integral in Eq. (1) by a sum,

$$\int_0^x \phi(z)K(x, z)\, dz = \sum_{\nu=0}^{\nu=n} A(n, \nu)\phi(\nu h) \tag{2}$$

where the coefficients $A(n, \nu)$ are determined by the kernel $K(x, z)$ and the value of h. Special formulae for the values of $A(n, \nu)$ are given below. Upon substituting Eq. (2) in Eq. (1) and solving for $\phi(nh)$, we obtain the recurrence formula

$$\phi(nh) = \frac{f(nh) - g(nh) \sum_{\nu=0}^{\nu=n-1} A(n, \nu)\phi(\nu h)}{1 + g(nh)A(n, n)} \tag{3}$$

According to Carson[2] and Collatz[4] the product $\phi(z)K(x, z)$ may be approximated by a linear function in each interval. Thus

$$\left.\begin{aligned} A(n, 0) &= \tfrac{1}{2}hK(nh, 0) \\ A(n, \nu) &= hK(nh, \nu h) \quad \text{for} \quad 1 \leqq \nu \leqq n - 1 \\ A(n, n) &= \tfrac{1}{2}hK(nh, nh). \end{aligned}\right\} \tag{4}$$

Approximation of the product $\phi(z)K(x, z)$ by a quadratic expression in each two consecutive intervals has been suggested by Collatz[4]; cf. also Section 3, Eqs. (32) to (39).

The approximation in Eq. (4) does not apply if the kernel $K(x, z)$ tends to infinity at $z = x$.

Huber[3] approximates the unknown function $\phi(z)$ by a linear expression in each interval,

$$\phi(z) = \phi(\nu h) + (\zeta/h)[\phi(\nu h + h) - \phi(\nu h)] \quad \text{if} \quad \nu h \leqq z \leqq \nu h + h \tag{5}$$

where

$$\zeta = z - \nu h. \tag{6}$$

This paper is based in part upon work sponsord by the Atomic Energy Commission under Contract AT-30-Gen-1002.

Hence

$$\int_{\nu h}^{\nu h+h} \phi(z)K(nh, z)\,dz = \phi(\nu h)[i_0(n, \nu) - i_1(n, \nu)] + \phi(\nu h + h)i_1(n, \nu) \quad (7)$$

where

$$i_0(n, \nu) = \int_0^h K(nh, \nu h + \zeta)\,d\zeta \quad (8)$$

$$i_1(n, \nu) = \int_0^h (\zeta/h)K(nh, \nu h + \zeta)\,d\zeta. \quad (9)$$

Adding up expressions in Eq. (7), one obtains the general form of Eq. (2) with the following special values of the coefficients $A(n, \nu)$

$$A(n, 0) = i_0(n, 0) - i_1(n, 0) \quad (10)$$

$$A(n, \nu) = i_0(n, \nu) + i_1(n, \nu - 1) - i_1(n, \nu) \quad \text{if} \quad 1 \leqq \nu \leqq n - 1 \quad (11)$$

$$A(n, n) = i_1(n, \nu - 1). \quad (12)$$

To account for deviations of the function $\phi(x)$ from linearity, Huber has worked out a special procedure by means of which corrections for the values obtained from Eqs. (3) and (8) to (12) are calculated.

Huber's method is especially helpful if the kernel tends to infinity at $z = x$. Since this case is particularly important in problems of physics and engineering, approximation of $\phi(x)$ by a quadratic expression is considered in the following in order to obtain a higher accuracy than with the aid of a linear approximation.

2. Solution of an Integral Equation If the Unknown Function Is Approximated by a Quadratic Function. If a quadratic expression for $\phi(x)$ in each two consecutive intervals is assumed, we have the interpolation formula

$$\begin{aligned}\phi(z) = \phi(\nu h + \zeta) &= \phi(\nu h) + (\zeta/h)[\phi(\nu h + h) - \phi(\nu h)] \\ &\quad - \tfrac{1}{2}[\zeta(h - \zeta)/h^2][\phi(\nu h + 2h) - 2\phi(\nu h + h) + \phi(\nu h)] \\ &= \phi(\nu h) + (\zeta/h)[2\phi(\nu h + h) - \tfrac{3}{2}\phi(\nu h) - \tfrac{1}{2}\phi(\nu h + 2h)] \quad (13) \\ &\quad + \tfrac{1}{2}(\zeta/h)^2[\phi(\nu h + 2h) - 2\phi(\nu h + h) + \phi(\nu h)] \\ &\qquad\qquad \text{for } \nu h \leqq z \leqq \nu h + 2h.\end{aligned}$$

Hence

$$\begin{aligned}\int_{\nu h}^{\nu h+2h} \phi(z)K(x, z)\,dz &= \int_0^{2h} \phi(\nu h + \zeta)K(nh, \nu h + \zeta)\,d\zeta \\ &= \phi(\nu h)[j_0(n, \nu) - \tfrac{3}{2}j_1(n, \nu) + \tfrac{1}{2}j_2(n, \nu)] \quad (14) \\ &\quad + \phi(\nu h + h)[2j_1(n, \nu) - j_2(n, \nu)] \\ &\quad + \phi(\nu h + 2h)[-\tfrac{1}{2}j_1(n, \nu) + \tfrac{1}{2}j_2(n, \nu)]\end{aligned}$$

where

$$j_0(n, \nu) = \int_0^{2h} K(nh, \nu h + \zeta)\, d\zeta \tag{15}$$

$$j_1(n, \nu) = \int_0^{2h} (\zeta/h) K(nh, \nu h + \zeta)\, d\zeta \tag{16}$$

$$j_2(n, \nu) = \int_0^{2h} (\zeta/h)^2 K(nh, \nu h + \zeta)\, d\zeta \tag{17}$$

Use of Eq. (14) for the evaluation of the integral in Eq. (1) requires an even number of intervals. We may therefore calculate $\phi(\frac{1}{2}h)$ separately, e.g., with the help of the Liouville-Neumann series. If n is even, we disregard the value of $\phi(\frac{1}{2}h)$ and add up expressions in Eq. (14) for $\nu = 0, 2, \cdots, (n-2)$. If n is uneven, we add up expressions in Eq. (14) for $\nu = 1, 3, \cdots, (n-2)$ and evaluate the integral $\int \phi(z)K(x, z)\, dz$ from $z = 0$ to $z = h$ with the aid of Eq. (14) and $h' = \frac{1}{2}h$,

$$\begin{aligned}\int_0^h \phi(z)K(x, z)\, dz &= \int_0^h \phi(\zeta)K(nh, \zeta)\, d\zeta \\ &= \phi(0)[i_0(n, 0) - 3i_1(n, 0) + 2i_2(n, 0)] \\ &\quad + \phi(\tfrac{1}{2}h)[4i, (n, 0) - 4i_2(n, 0)] \\ &\quad + \phi(h)[-i_1(n, 0) + 2i_2(n, 0)]\end{aligned} \tag{18}$$

where $i_0(n, 0)$ and $i_1(n, 0)$ are defined by Eqs. (8) and (9) and

$$i_2(n, 0) = \int_0^h (\zeta/h)^2 K(nh, \zeta)\, d\zeta. \tag{19}$$

Thus we obtain the general formula

$$\int_0^{nh} \phi(z)K(x, z)\, dz = \sum A(n, \nu)\phi(\nu h) \qquad \nu = 0, \tfrac{1}{2}, 1, 2, \cdots (n-1), n \tag{20}$$

where

$$A(n, 0) = j_0(n, 0) - \tfrac{3}{2}j_1(n, 0) + \tfrac{1}{2}j_2(n, 0) \text{ if } n \text{ is even} \tag{21}$$

$$A(n, 0) = i_0(n, 0) - 3i_1(n, 0) + 2i_2(n, 0) \text{ if } n \text{ is uneven} \tag{22}$$

$$A(n, \tfrac{1}{2}) = 0 \text{ if } n \text{ is even} \tag{23}$$

$$A(n, \tfrac{1}{2}) = 4i_1(n, 0) - 4i_2(n, 0) \text{ if } n \text{ is uneven} \tag{24}$$

$$A(n, 1) = 2j_1(n, 0) - j_2(n, 0) \text{ if } n \text{ is even} \tag{25}$$

$$\begin{aligned}A(n, 1) = -i_1(n, 0) + 2i_2(n, 0) + j_0(n, 1) - \tfrac{3}{2}j_1(n, 1) \\ + \tfrac{1}{2}j_2(n, 1) \text{ if } n \text{ is uneven and } n \geqq 3\end{aligned} \tag{26}$$

$$A(n, \nu) = 2j_1(n, \nu - 1) - j_2(n, \nu - 1) \text{ if } n - \nu \text{ is uneven and } 1 \leqq \nu \leqq n - 1 \tag{27}$$

$$A(n, \nu) = -\tfrac{1}{2}j_1(n, \nu - 2) + \tfrac{1}{2}j_2(n, \nu - 2) + j_0(n, \nu) - \tfrac{3}{2}j_1(n, \nu) + \tfrac{1}{2}j_2(n, \nu) \text{ if } n - \nu \text{ is even and } 2 \leqq \nu \leqq n - 2 \tag{28}$$

$$A(1, 1) = -i_1(1, 0) + 2i_2(1, 0) \tag{29}$$

$$A(n, n) = -\tfrac{1}{2}j_1(n, n - 2) + \tfrac{1}{2}j_2(n, n - 2) \text{ if } n \geqq 2 \tag{30}$$

If the kernel tends to infinity at $z = x$ but elsewhere is a smooth function and the width of the interval is not exceedingly small, Eq. (14) may be used only for the last two intervals from $z = nh - 2h$ to $z = nh$, and integrals for $z \leqq nh - 2h$ may be calculated with the aid of Simpson's rule, i.e., by approximating the product $\phi(z)K(x, z)$ by quadratic expressions in each two consecutive intervals,

$$\int_{\nu h}^{\nu h + 2h} \phi(z)K(nh, z)\, dz = h[\tfrac{1}{3}\phi(\nu h)K(nh, \nu h) + \tfrac{4}{3}\phi(\nu h + h)K(nh, \nu h + h) + \tfrac{1}{3}\phi(\nu h + 2h)K(nh, \nu h + 2h)] \quad \text{if } \nu \leqq n - 4 \tag{31a}$$

$$\int_0^h \phi(z)K(nh, z)\, dz = h[\tfrac{1}{6}\phi(0)K(nh, 0) + \tfrac{2}{3}\phi(\tfrac{1}{2}h)K(nh, \tfrac{1}{2}h) + \tfrac{1}{6}\phi(h)K(nh, h)] \quad \text{if } n \geqq 3 \tag{31b}$$

If Eqs. (31a) and (31b) are used in conjunction with Eq. (14) for $\nu = n - 2$ and Eq. (18) for $n = 1$, then Eq. (23) for $A(n, \tfrac{1}{2})$ if n is even, Eq. (29) for $A(1, 1)$, and Eq. (30) remain unchanged. Values of $A(1, 0)$, $A(2, 0)$, $A(1, \tfrac{1}{2})$, and $A(n, n - 1)$ are to be calculated from Eqs. (22), (21), (24), and (27), respectively. In addition, the following equations hold

$$A(n, 0) = \tfrac{1}{3}hK(nh, 0) \text{ if } n \text{ is even and } n \geqq 4 \tag{32}$$

$$A(n, 0) = \tfrac{1}{6}hK(nh, 0) \text{ if } n \text{ is uneven and } n \geqq 3 \tag{33}$$

$$A(n, \tfrac{1}{2}) = \tfrac{2}{3}hK(nh, \tfrac{1}{2}h) \text{ if } n \text{ is uneven and } n \geqq 3 \tag{34}$$

$$A(3, 1) = \tfrac{1}{6}hK(3h, h) + j_0(3, 1) - \tfrac{3}{2}j_1(3, 1) + \tfrac{1}{2}j_2(3, 1) \tag{35}$$

$$A(n, 1) = \tfrac{1}{2}hK(nh, h) \text{ if } n \text{ is uneven and } n \geqq 5 \tag{36}$$

$$A(n, \nu) = \tfrac{4}{3}hK(nh, \nu h) \text{ if } n - \nu \text{ is uneven and } 1 \leqq \nu \leqq n - 3 \tag{37}$$

$$A(n, \nu) = \tfrac{2}{3}hK(nh, \nu h) \text{ if } n - \nu \text{ is even and } 2 \leqq \nu \leqq n - 4 \tag{38}$$

$$A(n, n - 2) = \tfrac{1}{3}hK(nh, nh - 2h) + j_0(n, n - 2) - \tfrac{3}{2}j_1(n, n - 2) + \tfrac{1}{2}j_2(n, n - 2) \text{ if } n \geqq 4. \tag{39}$$

Calculation of coefficients $A(n, \nu)$ from Eqs. (32) to (39) will in general require less labor than use of Eqs. (21) to (30) and therefore is preferable if applicable.

3. The Choice of the Width of Intervals. It is desirable that minor errors disappear automatically when the calculation is continued. Practically, we may demand that the amount of the error $\varepsilon(nh)$ in $\phi(nh)$ be less than half the amount of the error $\varepsilon(nh - h)$ in the preceding value $\phi(nh - h)$. From Eq. (3) it follows that

$$\varepsilon(nh) = -\frac{g(nh)A(n, n-1)}{1 + g(nh)A(n, n)}\,\varepsilon(nh - h). \tag{40}$$

Hence we obtain the condition

$$\tfrac{1}{2} \mid 1 + g(nh)A(n, n) \mid > \mid g(nh)A(n, n-1) \mid \quad \text{for } \mid \varepsilon(nh) \mid < \tfrac{1}{2} \mid \varepsilon(nh - h) \mid . \tag{41}$$

In general, it is possible to satisfy Eq. (41) by choosing a sufficiently small value of h. Difficulties may occur if the function $g(x)$ is not bounded, e.g., increases exponentially as in Eq. (90), and a quadratic approximation is used. Use of a linear approximation according to Huber[3] may be preferable for large values of x in Eq. (90).

4. Non-linear Integral Equations. The procedure outlined above may also be used to obtain a numerical solution of the non-linear integral equation

$$\phi(x) = f(x) - g(x)\int_0^x P[\phi(z), z]K(x, z)\,dz. \tag{42}$$

If $P[\phi(z), z]$ is approximated by a quadratic expression in each two consecutive intervals, we have instead of Eq. (2)

$$\int_0^x P[\phi(z), z]K(x, z)\,dz = \sum A(n, \nu)P[\phi(\nu h), \nu h] \qquad \nu = 0, \tfrac{1}{2}, 1, 2, \cdots, (n-1), n \tag{43}$$

where the coefficients are the same as in Eq. (20). Upon substituting Eq. (43) in Eq. (42) for $x = nh$, we have instead of Eq. (3) the non-linear equation

$$\phi(nh) + g(nh)A(n, n)P[\phi(nh), nh] = f(nh) - g(nh)\sum_{\nu=0,\frac{1}{2},1,2,\cdots,(n-1)} A(n, \nu)P[\phi(\nu h), \nu h] \tag{44}$$

with $\phi(nh)$ as unknown. Eq. (44) may be solved for $\phi(nh)$ numerically by conventional methods[5].

The labor required for the calculation of $\phi(nh)$ from Eq. (44) will in general be insignificant in comparison to the labor required for the calculation of coefficients $A(n, \nu)$ and the sums in Eq. (44). Non-linearity is therefore not a serious complication provided that an approximation of $P[\phi(z), z]$ by a quadratic expression in each two consecutive intervals is adequate.

5. Special equations if the kernel is a function of $(x - z)$ only. If the kernel is a function of $(x - z)$ only, we introduce

$$s = x - z \tag{45}$$

as the variable of integration. Thus Eq. (1) becomes

$$\phi(x) = f(x) - g(x) \int_0^x \phi(x - s)K(s)\, ds \tag{46}$$

The integral in Eq. (46) for $x = nh$ may be approximated by a sum

$$\int_0^{nh} \phi(x - s)K(s)\, ds = \sum_{\mu=0}^{\mu=n} A_\mu \phi(nh - \mu h) \tag{47}$$

where μ is a serial number and the coefficients A_μ are determined below. Substituting Eq. (47) in Eq. (46) for $x = nh$, we obtain the recurrence formula

$$\varphi(nh) = \frac{f(nh) - \phi g(nh) \sum_{\mu=1}^{\mu=n} A_\mu \phi(nh - \mu h)}{1 + g(nh)A_0} \tag{48}$$

We assume that $\phi(x - s)$ can be represented by a quadratic expression in each two consecutive intervals. Thus, letting

$$\sigma = s - \mu h, \tag{49}$$

we have the interpolation formula

$$\begin{aligned}
\phi(x - s) &= \phi(nh - \mu h - \sigma) \\
&= \phi(nh - \mu h) + (\sigma/h)[\phi(nh - \mu h - h) - \phi(nh - \mu h)] \\
&\qquad - \tfrac{1}{2}[\sigma(h - \sigma)/h^2][\phi(nh - \mu h - 2h) \\
&\qquad - 2\phi(nh - \mu h - h) + \phi(nh - \mu h)] \\
&= \phi(nh - \mu h) + (\sigma/h)[2\phi(nh - \mu h - h) \\
&\qquad - \tfrac{3}{2}\phi(nh - \mu h) - \tfrac{1}{2}\phi(nh - \mu h - 2h)] \\
&\qquad + \tfrac{1}{2}(\sigma/h)^2[\phi(nh - \mu h - 2h) - 2\phi(nh - \mu h - h) \\
&\qquad + \phi(nh - \mu h)] \qquad \text{if } \mu h \leqq s \leqq \mu h + 2h.
\end{aligned} \tag{50}$$

Hence

$$\begin{aligned}
\int_{\mu h}^{\mu h + 2h} \phi(x - s)K(s)\, ds &= \int_0^{2h} \phi(nh - \mu h - \sigma)K(\mu h + \sigma)\, d\sigma \\
&= \phi(nh - \mu h)[\alpha_0(\mu) - \tfrac{3}{2}\alpha_1(\mu) + \tfrac{1}{2}\alpha_2(\mu)] \\
&\qquad + \phi(nh - \mu h - h)[2\alpha_1(\mu) - \alpha_2(\mu)] \\
&\qquad + \phi(nh - \mu h - 2h)[-\tfrac{1}{2}\alpha_1(\mu) + \tfrac{1}{2}\alpha_2(\mu)]
\end{aligned} \tag{51}$$

where

$$\alpha_0(\mu) = \int_0^{2h} K(\mu h + \sigma)\, d\sigma \tag{52}$$

$$\alpha_1(\mu) = \int_0^{2h} (\sigma/h)K(\mu h + \sigma)\, d\sigma \tag{53}$$

$$\alpha_2(\mu) = \int_0^{2h} (\sigma/h)^2 K(\mu h + \sigma)\, d\sigma \tag{54}$$

To obtain the integrals in Eq. (52) to (54) in a most convenient way, we may calculate first the following integrals

$$R(\mu h) = \int_0^{\mu h} K(s)\, ds \tag{55}$$

$$S(\mu h) = \int_0^{\mu h} sK(s)\, ds \tag{56}$$

$$T(\mu h) = \int_0^{\mu h} s^2 K(s)\, ds. \tag{57}$$

Hence

$$\alpha_0(\mu) = R(\mu h + 2h) - R(\mu h) \tag{58}$$

$$\alpha_1(\mu) = h^{-1}[S(\mu h + 2h) - S(\mu h)] - \mu[R(\mu h + 2h) - R(\mu h)] \tag{59}$$

$$\begin{aligned}\alpha_2(\mu) = h^{-2}[T(\mu h + 2h) - T(\mu h)] - 2\mu h^{-1}[S(\mu h + 2h) - S(\mu h)] \\ + \mu^2[R(\mu h + 2h) - R(\mu h)].\end{aligned} \tag{60}$$

If n is even, the integral on the left-hand side of Eq. (47) is obtained directly as the sum of expressions in Eq. (51). If n is uneven, the integral from $s = 0$ to $s = nh - h$ is calculated in the same fashion and the integral from $s = nh - h$ to $s = h$ is calculated separately with the aid of Eq. (51) for $h' = \frac{1}{2}h$. Thus

$$\begin{aligned}\int_{nh-h}^{nh} \phi(x - s)K(s)\, ds &= \int_0^h \phi(h - \sigma)K(nh - h + \sigma)\, d\sigma \\ &= \phi(h)[\beta_0(n-1) - 3\beta_1(n-1) + 2\beta_2(n-1)] \\ &\quad + \phi(\tfrac{1}{2}h)[4\beta_1(n-1) - 4\beta_2(n-1)] \\ &\quad + \phi(0)[-\beta_1(n-1) + 2\beta_2(n-1)]\end{aligned} \tag{61}$$

where

$$\beta_0(n-1) = \int_0^h K(nh - h + \sigma)\, d\sigma = R(nh) - R(nh - h) \tag{62}$$

$$\begin{aligned}\beta_1(n-1) &= \int_0^h (\sigma/h)K(nh - h - \sigma)\, d\sigma \\ &= h^{-1}[S(nh) - S(nh - h)] - (n-1)[R(nh) - R(nh - h)]\end{aligned} \tag{63}$$

$$\begin{aligned}\beta_2(n-1) &= \int_0^h (\sigma/h)^2 K(nh - h - \sigma)\, d\sigma \\ &= h^{-2}[T(nh) - T(nh - h)] - 2(n-1)h^{-1}[S(nh) - S(nh - h)] \\ &\quad + (n-1)^2[R(nh) - R(nh - h)].\end{aligned} \tag{64}$$

We obtain the general formula

$$\int_0^{nh} \phi(x-s)K(s)\,ds = \sum_{\mu=0,1,2,\cdots(n-1),(n-\frac{1}{2}),n} A_\mu \phi(nh-\mu h) \tag{65}$$

where

$$A_0 = \beta_0(0) - 3\beta_1(0) + 2\beta_2(0) \text{ if } n = 1 \tag{66}$$

$$A_0 = \alpha_0(0) - \tfrac{3}{2}\alpha_1(0) + \tfrac{1}{2}\alpha_2(0) \text{ if } n \geqq 2 \tag{67}$$

$$A_\mu = 2\alpha_1(\mu-1) - \alpha_2(\mu-1) \text{ if } \mu \text{ is uneven and } 1 \leqq \mu \leqq n-1 \tag{68}$$

$$A_\mu = -\tfrac{1}{2}\alpha_1(\mu-2) + \tfrac{1}{2}\alpha_2(\mu-2) + \alpha_0(\mu) - \tfrac{3}{2}\alpha_1(\mu) + \tfrac{1}{2}\alpha_2(\mu) \text{ if } \mu \text{ is even and } 2 \leqq \mu \leqq n-2 \tag{69}$$

$$A_{n-1} = -\tfrac{1}{2}\alpha_1(n-3) + \tfrac{1}{2}\alpha_2(n-3) + \beta_0(n-1) - 3\beta_1(n-1) + 2\beta_2(n-1) \quad \text{if } n \text{ is uneven} \tag{70}$$

$$A_{n-\frac{1}{2}} = 0 \text{ if } n \text{ is even} \tag{71}$$

$$A_{n-\frac{1}{2}} = 4\beta_1(n-1) - 4\beta_2(n-1) \text{ if } n \text{ is uneven} \tag{72}$$

$$A_n = -\tfrac{1}{2}\alpha_1(n-2) + \tfrac{1}{2}\alpha_2(n-2) \text{ if } n \text{ is even} \tag{73}$$

$$A_n = -\beta_1(n-1) + 2\beta_2(n-1) \text{ if } n \text{ is uneven.} \tag{74}$$

If the kernel tends to infinity at $z = x$ but elsewhere is a smooth function and the width of the interval is not exceedingly small, Eq. (51) may be used only for the first two intervals from $s = 0$ to $s = 2h$, and integrals for $s > 2h$ may be calculated with the aid of Simpson's rule, Eqs. (31a) and (31b). Then Eqs. (66), (67), and (71) remain unchanged and the following equations are to be used

$$A_{\frac{1}{2}} = 4\beta_1(0) - 4\beta_2(0) \text{ if } n = 1 \tag{75}$$

$$A_1 = -\beta_1(0) + 2\beta_2(0) \text{ if } n = 1 \tag{76}$$

$$A_1 = 2\alpha_1(0) - \alpha_2(0) \text{ if } n \geqq 2 \tag{77}$$

$$A_2 = -\tfrac{1}{2}\alpha_1(0) + \tfrac{1}{2}\alpha_2(0) \text{ if } n = 2 \tag{78}$$

$$A_2 = -\tfrac{1}{2}\alpha_1(0) + \tfrac{1}{2}\alpha_2(0) + \tfrac{1}{6}hK(2h) \text{ if } n = 3 \tag{79}$$

$$A_2 = -\tfrac{1}{2}\alpha_1(0) + \tfrac{1}{2}\alpha_2(0) + \tfrac{1}{3}hK(2h) \text{ if } n \geqq 4 \tag{80}$$

$$A_\mu = \tfrac{2}{3}hK(\mu h) \text{ if } \mu \text{ is even and } 4 \geqq \mu \geqq n-2 \tag{81}$$

$$A_\mu = \tfrac{4}{3}hK(\mu h) \text{ if } \mu \text{ is uneven and } 3 \geqq \mu \geqq n-1 \tag{82}$$

$$A_{n-1} = \tfrac{1}{3}hK(nh-h) \text{ if } n \text{ is uneven and } n \geqq 5 \tag{83}$$

$$A_{n-\frac{1}{2}} = \tfrac{2}{3}hK(nh-\tfrac{1}{2}h) \text{ if } n \text{ is uneven and } n \geqq 3 \tag{84}$$

$$A_n = \tfrac{1}{3}hK(nh) \text{ if } n \text{ is even and } n \geqq 4 \tag{85}$$

$$A_n = \tfrac{1}{6}hK(nh) \text{ if } n \text{ is uneven and } n \geqq 3 \tag{86}$$

6. Application to Heat Conduction Problems. The temperature T in a semi-infinite solid with uniform initial temperature T_0 at time $t = 0$ and a heat flux $q(t)$ as a given function of time from the surface into the interior of the solid is according to Carslaw and Jaeger[6]

$$T(\xi, t) = T_0 + [a^{\frac{1}{2}}/(\pi^{\frac{1}{2}}k)] \int_0^t q(t - \tau) \exp(-\xi^2/4a\tau)\tau^{-\frac{1}{2}}\, d\tau \tag{87}$$

where ξ is the distance from the surface, a is the thermal diffusivity of the solid, k its heat conductivity, and τ is an auxiliary variable with the dimension of time.

In practical cases, the heat flux is frequently not directly given but is known to be equal to the product of a heat transfer coefficient H and the difference of the temperature T_a of an ambient fluid and the surface temperature T_s, plus terms accounting for heat transmission due to radiation. The values of T_a and H are supposed to be given functions of time t. Then $q(t)$ has the general form

$$q(t) = H(t)[T_a(t) - T_s(t)] + r_1[T_a'(t)] - r_2[T_s(t)] \tag{88}$$

where $r_1(T_a')$ is the heat received by radiation per unit surface area per unit time from an another solid of temperature T_a', and $r_2(T_s)$ is the heat released by radiation per unit surface area per unit time.

Substituting Eq. (88) in Eq. (87) for $\xi = 0$ and using the auxiliary variable $\vartheta = t - \tau$, we obtain the following Volterra integral equation for the surface temperature T_s as unknown function of time t

$$T_s(t) = T_0 + [a^{\frac{1}{2}}/(\pi^{\frac{1}{2}}k)] \int_0^t \{H(\vartheta)[T_a(\vartheta) - T_s(\vartheta)] + r_1[T_a'(\vartheta)] - r_2[T_s(\vartheta)]\} \cdot (t - \vartheta)^{-\frac{1}{2}}\, d\vartheta \tag{89}$$

which has the general form of Eq. (42).

For instance, Eq. (89) may be used in order to calculate the surface temperature of a meteor traversing the atmosphere of the earth. At high altitudes, aerodynamic heating is negligible since the density of the ambient air is extremely low. With increasing time the heat transfer coefficient H increases quasi-exponentially. The unknown expression in braces as a function of ϑ under the integral in Eq. (89) may therefore be approximated by a quadratic expression in each two consecutive intervals.

If the heat transfer coefficient H is independent of temperature and heat transmission due to radiation is negligible, i.e., the terms r_1 and r_2 are disregarded, Eq. (89) may be transformed into a linear integral equation as stated in Eq. (1) or (46).

A similar diffusion problem is encountered in polarography. With the aid of the procedure outlined in Section 5, Delahay[7] has solved the integral equation

$$\phi(x) = e^{x-c} - e^{x-c} \int_0^x \phi(z)(x - z)^{-\frac{1}{2}}\, dz \tag{90}$$

where c is a constant parameter.

7. Special Procedure If the Unknown Function Cannot be Approximated by a Quadratic Expression. If the heat flux in Eq. (87) is finite from $t = 0$ on, the surface temperature initially increases proportionally to the square root of time Thus a linear or quadratic approximation for $T_s(t)$ at $t \cong 0$ is inadequate and the procedure outlined above fails.

This difficulty may be overcome by using $t^{\frac{1}{2}}$ and $\vartheta^{\frac{1}{2}}$ as the independent variable and the variable of integration, respectively. Then the kernel assumes the form $2z/(x^2 - z^2)^{\frac{1}{2}}$, which is not a function of $(x - z)$, and thus the computation of the coefficients $A(n, \nu)$ becomes rather cumbersome.

Another approach seems more profitable. If the first derivative of the unknown function $\phi(x)$ tends to infinity at $x = 0$, we may set $\phi(x)$ equal to the sum of an appropriately chosen function $\psi(x)$ accounting for the singularity at $x = 0$ and a function $\omega(x)$, which has to be determined by stepwise calculation,

$$\phi(x) = \psi(x) + \omega(x). \tag{91}$$

Substituting Eq. (91) in Eq. (1) with $K(x, z) = K(x - z)$, we obtain

$$\omega(x) = f^*(x) - g(x) \int_0^x \omega(x)K(x - z)\, dz \tag{92}$$

where

$$f^*(x) = f(x) - \psi(x) - g(x) \int_0^x \psi(z)K(x - z)\, dz. \tag{93}$$

The function $\psi(x)$ is supposed to give an adequate approximation for $\phi(x)$ at $x \cong 0$ so that at least the first derivative of $\omega(x)$ at $x = 0$ is finite, but otherwise the function $\psi(x)$ may be chosen at will. If possible the function $\psi(x)$ should be chosen so that a closed solution for the integral in Eq. (93) is obtainable. In addition, $\psi(x)$ should be chosen so that $\psi(x)$ remains finite for large values of x, or vanishes for $x \to \infty$.

To find $\psi(x)$, we may use the first two terms of the Liouville-Neumann series. For Eq. (1) with $K(x, z) = (x - z)^{-\frac{1}{2}}$ we have

$$\phi(x) \cong f(x) - g(x) \int_0^x f(0)K(x - z)\, dz = f(x) - 2x^{\frac{1}{2}}f(0)g(x) \qquad \text{for } x \cong 0. \tag{94}$$

Since the kernel $(x - z)^{-\frac{1}{2}}$ has a peak for $z = x$, we may also use an approximation suggested by Wagner[3] for Fredholm integral equations but likewise applicable to Volterra integral equations. Thus

$$\varphi(x) \cong \frac{f(x)}{1 + g(x) \int_0^x K(x, z)\, dz} = \frac{f(x)}{1 + 2x^{\frac{1}{2}}g(x)}. \tag{95}$$

8. Illustrative Example. For illustration, we shall solve Eq. (89) with constant values of T_s and H and vanishing values of r_1 and r_2. Introducing into Eq. (89) the dimensionless groups

$$x = H^2at/(\pi k^2) \tag{96}$$

$$z = H^2 a\vartheta/(\pi k^2) \tag{97}$$

$$\phi(x) = (T_a - T_s(t)/(T_a - T_0) \tag{98}$$

we obtain

$$\phi(x) = 1 - \int_0^x \phi(z)(x - z)^{-\frac{1}{2}}\, dz. \tag{99}$$

Since T_s necessarily tends to T_a for $t \to \infty$ and thus $\phi(x)$ vanishes for $x \to \infty$, we shall not use the right-hand member of Eq. (94) for $\psi = (x)$ since it tends to infinity for $x \to \infty$ if $g(x)$ remains finite. Instead, we may use Eq. (95) and therefore let

$$\psi = (1 + 2x^{\frac{1}{2}})^{-1}. \tag{100}$$

Substituting Eqs. (91) and (100) in Eq. (99), we obtain

$$\omega(x) = f^*(x) - \int_0^x \omega(x)(x - z)^{-\frac{1}{2}}\, dz \tag{101}$$

where

$$f^*(x) = 1 - \frac{1}{1 + 2x^{\frac{1}{2}}} - \int_0^x \frac{dz}{(1 + 2x^{\frac{1}{2}})(x - z)^{\frac{1}{2}}}. \tag{102}$$

Evaluation of the integral, I^*, in Eq. (102) gives

$$\left.\begin{aligned} I^* &= \tfrac{1}{2}\pi - (1 - 4x)^{\frac{1}{2}} \cos^{-1}(2x^{\frac{1}{2}}) \text{ if } x < \tfrac{1}{4} \\ I^* &= \tfrac{1}{2}\pi \text{ if } x = \tfrac{1}{4} \\ I^* &= \tfrac{1}{2}\pi - (4x - 1)^{\frac{1}{2}} \cosh^{-1}(2x^{\frac{1}{2}}) \text{ if } x > \tfrac{1}{4} \end{aligned}\right\} \tag{103}$$

The unknown function $\omega(x)$ determined by Eq. (101) may be computed with the aid of Eq. (48) and coefficients A_μ according to Eqs. (65) to (74) or (75) to (86). Then the function $\phi(x)$ is obtained from Eqs. (91) and (100) as

$$\phi(x) = (1 + 2x^{\frac{1}{2}})^{-1} + \omega(x). \tag{104}$$

Two different intervals, $h = 0.2$ and $h = 0.1$, were used. Values of $\phi(\frac{1}{2}h)$ were obtained with the aid of the Liouville-Neumann series. Values of $\phi(x)$ for $x = h, 2h, \cdots$ obtained by step-wise calculation agree satisfactorily with the values of the analytical solution of Carslaw and Jaeger[9]

$$\phi(x) = 1 - [1 - \operatorname{erf}(\pi^{\frac{1}{2}}x^{\frac{1}{2}})] \exp(\pi x) \tag{105}$$

as is shown in Table I.

9. Modified Procedure If a Stieltjes Integrimeter is Available. The evaluation of sums in Eqs. (3) or (48) is rather cumbersome if the number of intervals becomes large. In the following it is assumed that the kernel is a function of $(x - z)$. Calculating first

$$R(s) = \int_0^s K(\sigma)\, d\sigma, \tag{106}$$

TABLE I

Solution of integral equation (99)

$$\phi(x) = 1 - \int_0^x \phi(z)(x - z)^{-\frac{1}{2}}\, dz$$

x	$\phi(x)$ obtained by stepwise calculation		$\phi(x)$ exact solution according to Eq. (105)
	$h = 0.2$	$h = 0.1$	
0.0	1.0000	1.0000	1.0000
0.05	—	0.6730	0.6730
0.10	0.5859	0.5860	0.5859
0.2	0.4916	0.4916	0.4917
0.3	—	0.4357	0.4358
0.4	0.3965	0.3966	0.3967
0.5	—	0.3672	0.3671
0.6	0.3437	0.3436	0.3437
0.7	—	0.3244	0.3244
0.8	0.3079	0.3081	0.3082
0.9	—	0.2942	0.2942
1.0	0.2821	0.2820	0.2821

we obtain

$$\int_{2h}^{nh} \phi(nh - s)K(s)\, ds = \int_{s=2h}^{s=nh} \phi(nh - s)\, dR(s) \tag{107}$$

with the aid of the Stieltjes integrimeter.[10]. Substituting Eqs. (107) and (51) for $\mu = 0$ in Eq. (46), we obtain in view of Eqs. (67) and (68)

$$\phi(nh) = \frac{f(nh) - g(nh)\left[A_1\phi(nh - h) + A_2^*\phi(nh - 2h) + \int_{s=2h}^{s=nh} \phi(nh - s)\, dR(s)\right]}{1 + g(nh)A_0} \tag{108}$$

where according to Eq. (51)

$$A_2^* = -\tfrac{1}{2}\alpha_1(0) + \tfrac{1}{2}\alpha_2(0) \tag{109}$$

and A_0 and A_1 are given by Eqs. (67) and (68).

REFERENCES

1. G. Prasad, Proc. Edinburgh Math. Soc. **42**, 46 (1924).
2. J. R. Carson, Electrical Circuit Theory and the Operational Calculus, p. 145 (McGraw-Hill Book Company, Inc., New York, 1926).
3. A. Huber, Monatshefte f. Mathematik und Physik **47**, 240 (1939).
4. L. Collatz, Numerische Behandlung von Differentialgleichungen, pp. 411–415. Springer, Berlin, 1951.
5. J. B. Scarborough, Numerical Mathematical Analysis, pp. 185–203. The John Hopkins Press, Baltimore, 1950.
6. H. S. Carslaw and J. C. Jaeger, Conduction of Heat in Solids, p. 57. At the Clarendon Press, Oxford, 1947.

7. P. Delahay, J. Am. Chem. Soc., 75, 1190 (1953).
8. C. Wagner, Journ. Math. Phys. 30, 23 (1951).
9. H. S. Carslaw and J. C. Jaeger, loc. cit., p. 248.
10. E. J. Nystroem, Zeits. angew. Mathematik Mechanik 14, 276 (1943); W. Meyer zur Capellen, Mathematische Instrumente, pp. 208-211. Akademische Verlagsgesellschaft, Leipzig. 1944.

Department of Metallurgy
Massachusetts Institute of Technology

(Received February 13, 1953)

A NOTE ON SETS OF MATRICES SIMULTANEOUSLY REDUCIBLE TO THE TRIANGULAR SKELETON

By J. C. Eaves

1. Introduction. Many writers have considered the simultaneous reduction of square matrices of a set to the triangular skeleton [1, 2, 3, 4, 5]. It is usually more convenient both in application and in theory to use partitioned matrices. It is the purpose of this paper to show that a necessary and sufficient condition that a set of $n \times n$ matrices be reducible simultaneously by the same similarity transformation to matrices having the triangular skeleton is that they be simultaneously reducible to a set of regularly partitioned matrices having completely triangular blocks. This property is particularly advantageous in applied work where the matrices are often of very large order [6, 7]. For example, a 600×600 matrix may be so partitioned that 300×300 block matrices or matrices of even smaller dimension may be considered. In this way the desired results may be computed on the present available computers.

2. Notation. In the first sections of this paper we shall deal with a matrix of restricted order $\frac{1}{2}n(n+1)$. Later this restriction is removed. It is convenient here to use a double subscript, double superscript notation on the elements of the matrices considered. The block matrix A, of order $\frac{1}{2}n(n+1)$, is partitioned so that the diagonal blocks $A_{11}, A_{22}, \cdots, A_{nn}$ are of orders $n, n-1, \cdots, 1$ respectively. Thus, the block matrix A_{ij} is of dimension $(n-i+1) \times (n-j+1)$. The elements of A_{ij} are denoted by

$$a_{\phi,\theta}^{n-i+1,n-j+1} (\phi = 1, 2, \cdots, n-i+1; \theta = 1, 2, \cdots, n-j+1).$$

The superscripts of an element indicate both the block in which the element is located and the row-column dimension of that block. The subscripts indicate the row-column position of the element within the block A_{ij}. The element $a_{\phi,\theta}^{n-\lambda,n-\mu}$ is in the ϕ-row, θ-column of the block $A_{\lambda+1,\mu+1}$ which is of dimension $(n-\lambda) \times (n-\mu)$. We shall deal with matrices having elements in an algebraically closed field.

3. A transformation matrix. Let the matrix T, of order $\frac{1}{2}n(n+1)$, be defined as follows:

$$T = (T_{kl}) = (t_{ij}^{n-p,n-q}) \quad (p, q = 0, 1, \cdots, n-1; i = 1, 2, \cdots, n-p;$$

$$j = 1, 2, \cdots, n-q; k, l = 1, 2, \cdots, n) \tag{1}$$

$$t_{q+1,p+1}^{n-p,n-q} = 1, \qquad t_{ij}^{n-p,n-q} = 0 \qquad (i,j) \neq (q+1, p+1).$$

Since $t_{q+1,p+1}^{n-p,n-q} = t_{p+1,q+1}^{n-q,n-p} = 1$, it is evident that T is symmetric and orthogonal. Thus, $T = T^I$. Also, T is a permutative matrix.

4. Completely triangular matrices. We extend the definitions of triangular matrices and strictly triangular matrices to include rectangular matrices.

Definition. If the $m \times n$ matrix A is bordered either on the right by columns, or below by rows, of zeros, until it becomes a square matrix U and if U is a triangular matrix then we shall say A is an $m \times n$ triangular matrix. If U is strictly triangular, then we shall call A an $m \times n$ strictly triangular matrix.

Definition. If $A = (A_{ij})$ is a regularly partitioned matrix having $m \times n (m \geqq n)$ triangular matrices as the block matrices in and above the main diagonal and having $p \times q (p \leqq q)$ strictly triangular matrices as the block matrices below the main diagonal, then we shall call A a completely triangular matrix.

We first consider the completely triangular matrix A of restricted order $\frac{1}{2}n(n+1)$, defined by

$$A = (A_{ij}) = (\mathfrak{a}_{\phi\theta}^{n-\lambda,n-\mu}) \qquad (\lambda, \mu = 0, 1, \cdots, n-1; \phi = 1, 2, \cdots, n-\lambda;$$

$$\theta = 1, 2, \cdots, n-\mu; i, j = 1, 2, \cdots, n) \tag{2}$$

$$\mathfrak{a}_{\phi\theta}^{n-\lambda,n-\mu} = 0 \quad \text{for} \quad \lambda \leqq \mu,\ \phi > \theta; \lambda > \mu,\ \phi \geqq \theta.$$

Then, using the transformation matrix T, defined in 3, we have

$$T^I A T = U = (u_{ij}^{n-p,n-q}), \tag{3}$$

where the element $u_{ij}^{n-p,n-q}$ of U is the element $\mathfrak{a}_{p+1,q+1}^{n-(i-1),n-(j-1)}$ of A. The conditions on the p, q, i, j, satisfying the restriction that $u_{ij}^{n-p,n-q}$ be an element below the main diagonal of U are: (a) $p = q$ and $i > j$, (b) $p > q$. However, these conditions, imposed upon $\mathfrak{a}_{p+1,q+1}^{n-(i-1),n-(j-1)} = \mathfrak{a}_{\phi\theta}^{n-\lambda,n-\mu}$, give

(i) $\phi = \theta$ and $\lambda > \mu$ for $p = q$ and $i > j$,

(ii) $\phi > \theta$ and $\lambda > \mu$ for $p > q$ and $i > j$,

(iii) $\phi > \theta$ and $\lambda \leqq \mu$ for $p > q$ and $i \leqq j$.

Results (i) and (ii) may be written as

(iv) $\phi \geqq \theta$ and $\lambda > \mu$.

Conditions (iii) and (iv) are precisely the restriction (2). Consequently $T^I A T$ is a triangular matrix.

By a similar argument, $T^I U T$ is a completely triangular matrix if U is any $\frac{1}{2}n(n+1) \times \frac{1}{2}n(n+1)$ triangular matrix. Thus we have:

THEOREM 1. *Let $\Sigma = \{A_i ; i = 1, 2, \cdots, r\}$ be a finite set of $\frac{1}{2}n(n+1) \times \frac{1}{2}n(n+1)$ matrices. A necessary and sufficient condition that the matrices of Σ be reducible simultaneously to triangular matrices is that they be reducible by the same similarity transformation to completely triangular matrices.*

Just as in the triangular matrix, the characteristic roots of a completely triangular matrix are the diagonal elements. For, from (3), the diagonal element $\mathfrak{a}_{p+1,p+1}^{n-(i-1),n-(i-1)}$ of A is the diagonal element $u_{i,i}^{n-p,n-p}$ of $T^I A T$. From this it

follows that the value of the determinant of a completely triangular matrix A is the product of the diagonal elements.

5. Removal of the order restrictions. We now remove the order restriction, imposed on (A_{ij}). Let the $r \times r$ matrix $M = (M_{ij})(i, j = 1, 2, \cdots, t)$ be any principal submatrix of (A_{ij}) and indicate the orders of the diagonal blocks $M_{11}, M_{22}, \cdots, M_{tt}$ by $r_1, r_2, \cdots, r_t$ respectively, $r_i \neq r_j$ for $i \neq j$. There is no loss in generality to require that $r_i > r_j$ for $i < j$ since, if this were not the case a similarity transformation exists which will bring about this arrangement of these blocks. We now border M in such a manner as to obtain the matrix $\operatorname{diag}\{M, I\}$ of order $\frac{1}{2}r_1(r_1 + 1)$. There exists a permutative block matrix P, such that

$$P^I \begin{pmatrix} M & 0 \\ 0 & I \end{pmatrix} P = N = (N_{ij}), \tag{4}$$

where the block N_{ii} is of order $(r_1 - i + 1)$. By Theorem 1 there exists a matrix T such that T^INT is a triangular matrix and the diagonal elements of N are the diagonal elements of T^INT. Furthermore, the diagonal elements of I in $\operatorname{diag}\{M, I\}$, being also diagonal elements of T^INT, are the only non-zero elements in the rows and columns which contain these unit elements. Thus there exists a permutative matrix Q such that $Q^IT^INTQ = \operatorname{diag}\{S, I\}$ where the S is an $r \times r$ triangular matrix whose elements are those of M, and whose diagonal elements are the diagonal elements of M. The matrix S can be obtained directly from M by a similarity transformation on M. For, the similarity transformation of $\operatorname{diag}\{M, I\}$ by the matrix PTQ affects the last $\{(\Sigma_{i=1}^{r} i) - r\}$ rows and columns as does the identity transformation on these rows and columns. It is also seen that PTQ is a permutative matrix since it is a product of elementary permutative matrices, the set of $n!$ such matrices forming a symmetric group and thus being closed with respect to the operation multiplication.

Now, let $X = (X_{ij})(i, j = 1, 2)$ where X_{22} is an $m \times m$ submatrix, be any permutative matrix which leaves the last m rows of every matrix Y unaltered under the multiplication XY. Consider the special matrix $Y = \operatorname{diag}\{Y_{11}, I_m\}$. Then

$$XY = \begin{pmatrix} X_{11}Y_{11} & X_{12} \\ X_{21}Y_{11} & X_{22}I_m \end{pmatrix} = \begin{pmatrix} X_{11}Y_{11} & X_{12} \\ 0 & I_m \end{pmatrix} \quad \text{and} \quad X_{12} = 0.$$

Thus the matrix PTQ has the skeleton $\operatorname{diag}\{R, I\}$ where R is a permutative matrix of order r. Thus R^IMR is a triangular matrix. Similarly, for a triangular matrix M then R^IMR is a completely triangular matrix. Thus Theorem 1 holds for this more general case.

By a similar argument and induction it is easy to show that there exists a permutative matrix P such that, for a completely triangular matrix A, then P^IAP is a triangular matrix.

6. Other cases. The utility of the above results lies not only in its application to strictly triangular matrices. It can be applied to certain blocks of block

matrices which are only "partially" or "almost" strictly triangular. As in illustration it will be more convenient to consider a pair of 10×10 matrices, with which the writer was confronted recently, rather than to use 100×100 matrices or matrices of dimension 400 or 600 about which we are beginning to hear in some recent research and applied work.

After a rather obvious similarity transformation and a second similarity transformation which was not so obviously applicable the two 10×10 matrices A and B were reduced to matrices having the following skeleton:

$$A = \left(\begin{array}{cccc|ccc|ccc}
\alpha_{00} & \alpha_{01} & \alpha_{02} & \alpha_{03} & \alpha_{04} & \alpha_{05} & \alpha_{06} & 0 & \alpha_{08} & \alpha_{09} \\
0 & \alpha_{11} & 0 & \alpha_{13} & 0 & \alpha_{15} & \alpha_{16} & 0 & 0 & \alpha_{19} \\
\hline
0 & \alpha_{21} & \alpha_{22} & \alpha_{23} & 0 & 0 & \alpha_{26} & 0 & 0 & \alpha_{29} \\
0 & 0 & 0 & \alpha_{33} & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
0 & \alpha_{41} & \alpha_{42} & \alpha_{43} & \alpha_{44} & \alpha_{45} & \alpha_{46} & 0 & \alpha_{48} & \alpha_{49} \\
0 & 0 & \alpha_{52} & \alpha_{53} & 0 & \alpha_{55} & \alpha_{56} & 0 & 0 & \alpha_{59} \\
0 & 0 & 0 & \alpha_{63} & 0 & 0 & \alpha_{66} & 0 & 0 & 0 \\
\hline
\alpha_{70} & \alpha_{71} & \alpha_{72} & \alpha_{73} & \alpha_{74} & \alpha_{75} & \alpha_{76} & \alpha_{77} & \alpha_{78} & \alpha_{79} \\
0 & \alpha_{81} & \alpha_{82} & \alpha_{83} & 0 & \alpha_{85} & \alpha_{86} & 0 & \alpha_{88} & \alpha_{89} \\
0 & 0 & 0 & \alpha_{93} & 0 & 0 & \alpha_{96} & 0 & 0 & \alpha_{99}
\end{array}\right)$$

The similarity transformation of A by P and of B by P, where $P = (\epsilon_8\epsilon_1\epsilon_5\epsilon_9\epsilon_6\epsilon_3\epsilon_2\epsilon_{10}\epsilon_7\epsilon_4)$ and where ϵ_i is the 1×10 matrix with 1 in the i^{th} position and zeros elsewhere, transforms A and B into the triangular block matrices (A_{ij}), (B_{ij}) such that A_{11}, A_{22}, B_{11} and B_{22} are 2×2 triangular matrices. A_{44} and B_{44} are 3×3 triangular matrices. The blocks A_{33} and B_{33} were the following:

$$A_{33} = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & -1 \\ 2 & 0 & -1 \end{pmatrix}, \qquad B_{33} = \begin{pmatrix} -1 & -2 & 0 \\ 0 & 1 & 2 \\ 1 & 0 & 1 \end{pmatrix}.$$

Thus, the original problem of finding the characteristic roots of a polynomial, $f(A, B)$, was reduced to a study of this function for the 3×3 matrices A_{33} and B_{33}.

It is clear that no permutative transformation exists which will transform either A_{33} or B_{33} into the triangular skeleton. Furthermore, since $A_{33}B_{33} - B_{33}A_{33}$ is nonsingular it is obvious that no transformation matrix P exists such that $P^IA_{33}P$ and $P^IB_{33}P$ are of triangular skeleton.

It is of interest to note that the 4×4 matrices $A = (\alpha_{ij})$, $B = (b_{ij})$ where $\alpha_{12} = b_{31} = \alpha \neq 0$ and $\alpha_{ij} = b_{ij} = 0$ for all other elements are not simultaneously reducible to triangular skeleton.

REFERENCES

1. Williamson, J., The simultaneous reduction of two matrices to triangle form. Amer. J. Math. **57**: 281–293, 1935.
2. Paige, L. J., Simultaneous reduction of matrices to triangular form (Abstract). Bull. Amer. Math. Soc. **57**: 369, 1951.
3. Drazin, M. P., Dungey, J. W. and Gruenberg, K. W., Some theorems on commutative matrices. Jour. London Math. Soc. **26**, 221–228, 1951.
4. Eaves, J. C., On sets of matrices having a delayed commutativity property. J. of the Elisha Mitchell Scientific Soc. **68**: 46–54, 1952.
5. McCoy, N. H., On the characteristic roots of matric polynomials. Bull. Amer. Math. Soc. **42**: 592–600, 1936.
6. Householder, A. S., The relation of mathematics to the physical sciences. Texas J. Sc. **2**: 196–200, 1950.
7. U. S. Army, Matrix algebra. Ballistic Basic Research, Army Project 599-16-005 (TB3-0001), Technical Objective BR-1.

Alabama Polytechnic Institute

(Received March 31, 1952)

NOTE ON AN EQUI-MOMENTAL COMPLEX OF A RIGID BODY

By Hari Das Bagchi

Introduction. The main object of the present investigation is the study of certain novel properties of an "equi-momental complex" of a rigid body, *i.e.*, of the complex formed by the assemblage of ∞^3 lines (inside a body), the moment of inertia round any of which has a *fixed* pre-assigned (scalar) value I. This subject was originally initiated by Jung[1] and was latterly written upon by Painvin[2] and Demoulin.[3, 4]

It is not too much to say that the main bulk of this paper is devoted to a discussion of the surface-loci and curve-loci, traced out by a (variable) point 'A', when its associated cone Σ_A is subjected to different varieties of geometrical conditions.

On the whole the paper is believed to embody some amount of *original* matter, although there are casual references to *known* results.

Art. 1. The moment of inertia (*M.I.*) of a rigid body with respect to a line through the mass centre O in the direction $\mathbf{u}$, $\mathbf{u}\cdot\mathbf{u} = 1$, can be given by the formula $I_0(\mathbf{u}) = \mathbf{u}\cdot\varphi_0(\mathbf{u})$, where

$$\varphi_0(\mathbf{u}) = \Sigma m_i \mathbf{r}_i \times (\mathbf{u} \times \mathbf{r}_i),$$

(or the corresponding integral) $\mathbf{r}_i$ being the radius vector of the mass point m_i.[5] The *M.I.* with respect to an *arbitrary* line in the direction $\mathbf{u}$ through a point A of radius vector $\mathbf{a}$ is

$$I_A(\mathbf{u}) = \mathbf{u}\cdot\varphi_0(\mathbf{u}) + M(\mathbf{a} \times \mathbf{u})\cdot(\mathbf{a} \times \mathbf{u}) = \mathbf{u}\cdot\varphi_A(\mathbf{u}),$$

where

$$\varphi_A(\mathbf{u}) = \varphi_0(\mathbf{u}) + M(\mathbf{a} \times \mathbf{u}) \times \mathbf{a} \qquad \text{and} \qquad M = \Sigma m_i .$$

The totality of lines with respect to which the *M.I.* is equal to a *specified* (scalar) constant I constitutes a quadratic complex Ξ for which

$$\mathbf{u}\cdot\varphi_0(\mathbf{u}) + M(\mathbf{a} \times \mathbf{u})\cdot(\mathbf{a} \times \mathbf{u}) - I = 0. \tag{1}$$

If Σ_A is the "cone of the complex"[6] attached to A, its scalar equation is, according to (1), quadratic and of the form

$$\mathbf{r}\cdot\varphi(\mathbf{r}) + M(\mathbf{a} \times \mathbf{r})\cdot(\mathbf{a} \times \mathbf{r}) - I\mathbf{r}\cdot\mathbf{r} = 0, \tag{2}$$

[1] G. Jung: *Enzyklopädie der Mathem. Wissenschaften*, **4**, 9, p. 322.

[2] Painvin: *Nouv. Annales de Mathém.*, 1072, p. 49.

[3] A. Demoulin: *Bulletin Société Mathém. France*, **19**, 1891, p. 131.

[4] See also H. D. Bagchi: 'Normal Complex of a rigid body', *Bull. Cal. Math. Soc.*, **31**, No. 2, (1939).

[5] See e.g. C. J. Coe, Theoretical mechanics (N. Y. MacMillan, 1948), Ch. XI.

[6] Σ_A is also called the "equimomental cone at the point A." See Routh, Elementary Rigid Dynamics (1930), pp. 21–22, Ex. 32.

where $\mathbf{r}$ is the radius vector from A (*not* from 0). If then

$$\omega_A(\mathbf{u}) = \varphi_0(\mathbf{u}) + M\{(\mathbf{a} \times \mathbf{u}) \times \mathbf{a}\} - I\mathbf{u} = \varphi_A(\mathbf{u}) - I\mathbf{u}, \tag{3}$$

then the equation of Σ_A can be written either in the form

$$\mathbf{r}\cdot\omega_A(\mathbf{r}) = 0, \quad \text{or} \quad \mathbf{r}\cdot\varphi_A(\mathbf{r}) - I\mathbf{r}\cdot\mathbf{r} = 0. \tag{4}$$

Art. 2. Let $\mathbf{i}_1, \mathbf{i}_2, \mathbf{i}_3$ be mutually perpendicular unit vectors in the principal directions of inertia at C. Then there exist three scalars g_1, g_2, g_3 such that

$$\varphi_0(\mathbf{i}_1) = g_1\mathbf{i}_1, \qquad \varphi_0(\mathbf{i}_2) = g_2(\mathbf{i}_2), \qquad \varphi_0(\mathbf{i}_3) = g_3\mathbf{i}_3.$$

Then, if $\mathbf{u} = u_1\mathbf{i}_1 + u_2\mathbf{i}_2 + u_3\mathbf{i}_3$, $\mathbf{a} = a_1\mathbf{i}_1 + a_2\mathbf{i}_2 + a_3\mathbf{i}_3$, we find from (3)

$$\omega_A(\mathbf{i}_1) = p_1\mathbf{i}_1 - Ma_1\mathbf{a}, \qquad \omega_A(\mathbf{i}_2) = p_2\mathbf{i}_2 - Ma_2\mathbf{a}, \qquad \omega_A(\mathbf{i}_3) = p_3\mathbf{i}_3 - Ma_3\mathbf{a},$$

where $p_1 = g_1 + M\mathbf{a}\cdot\mathbf{a} - I$, $p_2 = g_2 + M\mathbf{a}\cdot\mathbf{a} - I$, $p_3 = g_3 + M\mathbf{a}\cdot\mathbf{a} - I$. We thus obtain for the three scalar invariants of the vector function $\omega_A(\mathbf{r})$ the following expressions[7]

$$m_A = (\omega_A(\mathbf{i}_1), \omega_A(\mathbf{i}_2), \omega_A(\mathbf{i}_3)) = p_1p_2p_3 - Ma_1^2p_2p_3 - Ma_2^2p_3p_1 - Ma_3^2p_1p_2 \tag{5}$$

$$\begin{aligned} m'_A &= (\mathbf{i}_1\omega_A(\mathbf{i}_2)\omega_A(\mathbf{i}_3)) + (\mathbf{i}_2\omega_A(\mathbf{i}_3)\omega_A(\mathbf{i}_1)) + (\mathbf{i}_3\omega_A(\mathbf{i}_1)\omega_A(\mathbf{i}_2)) \\ &= (p_1p_2 + p_2p_3 + p_3p_1) - Ma^2(p_{12} + p_3) - Ma_2^2(p_3 + p_1) - Ma_3^2(p_1 + p_2) \end{aligned} \tag{6}$$

$$\begin{aligned} m''_A &= \mathbf{i}_1\cdot\omega_A(\mathbf{i}_1) + \mathbf{i}_2\cdot\omega_A(\mathbf{i}_2) + \mathbf{i}_3\cdot\omega_A(\mathbf{i}_3) \\ &= (p_1 + p_2 + p_3) - M(a_1^2 + a_2^2 + a_3^2) \\ &= (g_1 + g_2 + g_3) - 3I + 2M(a_1^2 + a_2^2 + a_3^2). \end{aligned} \tag{7}$$

Art. 3. We now investigate four specific loci traced out by a *variable* point A (inside the body), when the cone Σ_A is made to satisfy certain *prescribed* conditions.

Case I. The quadratic cone Σ_A has a double line (is a pair of planes). Then $m_A = 0$. Hence, according to (5), substituting for $\mathbf{a}$ the generic vector $\mathbf{r} = x\mathbf{i} + y\mathbf{j} + 2\mathbf{k}$ we find for the locus

$$\frac{x^2}{[(g_1 - I)/M] + \mathbf{r}\cdot\mathbf{r}} + \frac{y^2}{[(g_2 - I)/M] + \mathbf{r}\cdot\mathbf{r}} + \frac{z^2}{[(g_1 - I)/M] + \mathbf{r}\cdot\mathbf{r}} = 1 \tag{8}$$

We thus corroborate, by an *alternative* method, a result *previously known*, viz.

Prop. I. The singular surface of the quadratic complex Ξ is the quartic surface Ω defined by Eq. (8).

This quartic surface is identical with Fresnel's wave surface.[8]

[7] See e.g. L. Brand, Vector and tensor analysis (N. Y. Wiley), 1947, p. 149. We have written $(\mathbf{abc}) = (\mathbf{a}, \mathbf{b}, \mathbf{c})$ for $(\mathbf{a} \times \mathbf{b})\cdot\mathbf{c}$.

[8] See Routh, loc. cit. p. 43, art. 65.

Case II. The cone Σ_A possesses triads of perpendicular generators. Then $m''_A = 0$. Hence, according to (7), we find for the locus

$$x^2 + y^2 + z^2 = \frac{3I - (g_1 + g_2 + g_3)}{2M}. \tag{9}$$

We thus arrive at the following proposition:

Prop. II. The locus of a point A (inside the body), for which the associated cone Σ_A admits of triads of perpendicular generators, is the sphere π given by (9).

This sphere has its centre at the mass centre of the body and is real when $3I > g_1 + g_2 + g_3$.

Case III. The cone Σ_A possesses triads of perpendicular tangent planes. Then $m'_A = 0$. Hence, according to (6), we find for the locus

$$(x^2 + y^2 + z^2)^2 - lx^2 - my^2 - nz^2 + D = 0, \tag{10}$$

where

$$l = \frac{4I - (2g_1 + g_2 + g_3)}{M}, \qquad m = \frac{4I - (g_1 + 2g_2 + g_3)}{M},$$

$$n = \frac{4I - (g_1 + g_2 + 2g_3)}{M}$$

$$D = \frac{(g_1 - I)(g_2 - I) + (g_2 - I)(g_3 - I) + (g_3 - I)(g_1 - I)}{M^2}.$$

Prop. III. The locus of a point A (inside the body), for which the associated cone Σ_A admits of triads of perpendicular tangent planes, is the surface Λ of the fourth degree, represented by Eq. (10).

The very form of (10) suggests that the surface Λ is a *cyclide of the central type*,[9] having O for its center. To find other particulars about Λ we set $D' = k^2$, $l' = l + 2k^2/A$, $m' = m + 2k^2/A$, $n' = n + 2k^2/A$, so as to re-write (10) in the equivalent form

$$(x^2 + y^2 + z^2 + k^2)^2 = A(l'^2x^2 + m'^2y^2 + n'^2z^2). \tag{11}$$

Comparing (11) with the "standard form" of a cyclide, as given in Art. 561 (a) of Salmon's "*Treatise on the Analytical Geometry of Three Dimensions*" (Vol. II, Rogers, 1915), we learn that the (central) cyclide admits of *only two*—and not five—distinct "focal quadrics"[10] viz.,

$$\frac{x^2}{l'^2} + \frac{y^2}{m'^2} + \frac{z^2}{n'^2} = 1 \quad \text{and} \quad \frac{x^2}{l'^2 - k^2} + \frac{y^2}{n'^2 - k^2} + \frac{z^2}{n'^2 - k^2} = 1,$$

[9] See H. D. Bagchi: (a) Note on Cyclides. *Journal of the Indian Mathematical Soc.*, Bangalore, India (1940). (b) Note on Cyclides and Hypercyclides. *Journal of the Asiatic Soc. of Bengal*, Calcutta (1949).

[10] It is worthy of note that whereas the *average* type of cyclide admits of *five* "focal quadrics" and *five* associated "spheres of inversion," the *specialized variety* of cyclide, *viz.* Λ admits of only *two* "focal quadrics and *two* associated "spheres of inversion." For obvious reasons, these two spheres cannot be both real.

the corresponding "spheres of inversion" being respectively

$$x^2 + y^2 + z^2 = k^2 \quad \text{and} \quad x^2 + y^2 + z^2 = -k^2.$$

Case IV. The cone Σ_A is a cone of revolution. Then the vector function $\omega_A(\mathbf{r})$ defined by (3) must have two *equal* roots. Now $\varphi_A(\mathbf{r})$ and $\omega_A(\mathbf{r})$ are coaxial and the three (scalar) roots of one differ from those of the other by I. The equality of two roots of $\omega_A(\mathbf{r})$ implies that of two roots of $\varphi_A(\mathbf{r})$ and vice versa, i.e. *the cone Σ_A will be right circular, whenever the momental ellipsoid at A is a spheroid and conversely*. In other words, Σ_A will be right circular when and only when two of the principal moments of inertia at A are equal. Hence recollecting that the two *reciprocal* quadrics at A, *viz.*, the momental ellipsoid and the ellipsoid of gyration at the point, are either both spheroidal or both non-spheroidal, and referring to Art. 62 (p. 41) of Routh (*loc. cit.*), we arrive at the following proposition.

Prop. IV. The elliptic and hyperbolic focal conics of the ellipsoid of gyration at its mass centre constitute the complete locus of a point A

a) *whose associated cone Σ_A is right circular, or (what is the same thing),*

b) *whose attached momental ellipsoid (or ellipsoid of gyration) is a spheroid.*

Art. 4. It now remains to give a finishing touch to the present investigation by taking into account certain subsidiary cases, arising from different combinations of special pairs of cases, dealt with in Art. 3. To be precise, we shall reckon with *three* such specific combinations, *viz.*:

Case V: Cases I and II *combined*;
Case VI: Cases II and IV *combined*;
and *Case VII*: Cases III and IV *combined*.

Case V—Amalgamating Cases I and II and remembering that a *degenerate* (quadric) cone, which possesses triads of perpendicular generators, is no else than a pair of *orthogonal* planes, we arrive at the following proposition:

Prof. V. The twisted curve, formed by the intersection of the "equi-momental" surface Ω, (given by (8), with the sphere (given by (9), represents the locus of a point A (inside the body), the associated cone at which consists of two perpendicular planes.

Case VI. Next amalgamating Cases II and IV and bearing in mind that a right circular cone, endowed with triads of perpendicular generators, must have $\tan^{-1}(\sqrt{2})$ for its semi-vertical angle, we arrive at the following proposition:

Prop. VI. Those points (inside a rigid body), the "associated" cones at which are right circular cones of the same semi-vertical angle $\tan^{-1}(\sqrt{2})$, *are designable geometrically as the points of intersection of the sphere π, (given by* (9), *with the elliptic and hyperbolic focal conics of the ellipsoid of gyration at the mass-centre (of the body).*

Case VII. Finally amalgamating Cases III and IV and recognising that a

right circular cone, possessing triads of perpendicular tangent planes, must have $\cot^{-1}(\sqrt{2})$ for its semi-vertical angle, we arrive at the following proposition:

Prop. VII. Those points (inside a rigid body) the "associated" cones whereat are right circular cones of the same semi-vertical angle $\cot^{-1}(\sqrt{2})$ *are definable as the points of intersection of the cyclide* Λ, *(given by* (10)*), with the elliptic and hyperbolic focal conics of the ellipsoid of gyration at the mass-centre.*

Inquisitive readers may widen the scope of enquiry and ascertain the locus of a point A for which Σ_A shall be a right circular cone of *any assigned* vertical angle. It is hardly necessary to remark that a combination of Cases II and III in the above context serves no useful purpose, for a quadric cone, endowed simultaneously with triads of perpendicular tangent planes, can never be real.

In conclusion, I beg to express my indebtedness to Professor D. J. Struik for his suggestions and helpful criticism.

CALCUTTA UNIVERSITY

(Received December 19, 1952)

BOUNDARY LAYER CONTROL BY DISTRIBUTED SURFACE SUCTION OR INJECTION. BI-PARAMETRIC GENERAL SOLUTION

By T. P. Torda

Introduction. In reference 1, a method was presented for the analysis of boundary layer control by distributed surface suction or fluid injection. This method was based on the von Kármán integral principle and used a polynomial approximation for the velocity profile in the boundary layer. New boundary conditions were introduced which allowed the evaluation of a physically significant boundary layer thickness and the variation of the required suction distribution necessary to maintain a prescribed laminar boundary layer with constant or varying thickness. In reference 1, only the momentum equation was used and, therefore, numerical methods (e.g. isoclines) were necessary for the final integration of the resulting differential equation. Although other investigators used an appropriate energy equation for the evaluation of the boundary layer variation along solid surfaces, in all previous work known to the author, the Prandtl momentum equation alone is used for the analysis of boundary layer control by distributed suction. The method described here employs the boundary layer momentum equation and the energy equation for the analysis of boundary layer control by distributed surface suction or fluid injection. Again, the von Kármán integral principle is used. The tangential velocity distribution in the boundary layer is approximated by an infinite series in powers of the normal coordinate with coefficients which are functions of the tangential coordinate. For flows around an airfoil, use of the terms containing the first four powers of the normal coordinate are sufficient. The two resulting simultaneous differential equations allow the determination of both the boundary layer thickness variation and the corresponding suction velocity distribution. However, in most cases of applications of boundary layer control the behavior of the boundary layer is prescribed (since the maintenance of a thin, laminar boundary layer over the whole profile is desired). Therefore, the two differential equations are combined and yield a two-parameter equation for the suction velocity distribution. International Business Machines computers were used for the evaluation of the resulting two-parameter polynomials and families of curves were plotted allowing the direct evaluation of the problem instead of the laborious numerical or graphical computations which were necessary previously. The boundary conditions which were introduced in reference 1 were employed. These are: The tangential velocity component is equal to the velocity of the potential flow at the edge of the boundary layer. Here, the normal derivative of the tangential velocity component is zero. It should be noted that these two boundary conditions are approximate only. Further, at the wall, the tangential velocity is either zero or has a finite magnitude (slip velocity) and the normal velocity component is equal to the suction (or injection) velocity. The boundary layer momentum equation and its derivatives with respect to the normal coordinate, evaluated at the wall,

yield additional boundary conditions necessary for the evaluation of the coefficients of the power series approximation of the tangential velocity component in the boundary layer.

Method of solution. Let s and n denote cartesian coordinates measured along the surface and perpendicular to it. The quantities u and v are the velocity components in the boundary layer in the directions of s and n, respectively, $U = U(s)$ is the velocity of the potential flow parallel to the surface, $p = p(s)$ is the pressure, ρ is the density, ν the kinematic viscosity, δ the boundary layer thickness and $\epsilon = \epsilon(s)$ is a function modifying the shear stress (due to slip at the wall). ϵ was taken as unity in the analysis. Subscript zero denotes conditions at the wall. Since $p(s)$ is the same in the potential flow and in the boundary layer, the boundary layer momentum equation and the energy equation may be written as:

$$u\,\partial u/\partial s + v\,\partial u/\partial n = UdU/ds + \nu\,\partial^2 u/\partial n^2 \tag{1}$$

$$u^2\,\partial u/\partial s + uv\,\partial u/\partial n = uUdU/ds + \nu u\,\partial^2 u/\partial n^2 \tag{2}$$

The approximation for the velocity component parallel to s in the boundary layer is taken as

$$u = U(an + bn^2 + cn^3 + dn^4) \tag{3}$$

where a, b, c, and d, are functions of s. These coefficients are evaluated from the boundary conditions such that

$$\left.\begin{aligned} a &= (24 + 6N + MN)/\delta D \quad b = 3(-3N + 4M)/\delta^2 D) \\ c &= (-3MN + 4M^2)/\delta^3 D) \cdot d = (-6 + 3N + 2MN - 6M - 3M^2)/\delta^4 D \\ D &= 18 + 6M - M^2 \end{aligned}\right\} \tag{4}$$

In the above

$$M = v_0\delta/\nu \text{ and } N = \delta^2(dU/ds)/\nu \tag{5}$$

Equations (1) and (2) were integrated (v at $y = \delta$ was calculated from the continuity equation) and use was made of expressions (3), (4), and (5). Two simultaneous total differential equations result (prime denotes d/ds). Introduction of the abbreviations

$$z \equiv \delta^2/\nu \qquad \sigma \equiv UU''/U'^2 \qquad \eta \equiv Uz' \tag{6}$$

allows the expression of the two simultaneous total differential equations as

$$Uv_0'(z)^{3/2}(\nu)^{-1/2} = A_1 + A_2\sigma + A_3\eta \tag{7}$$

$$Uv_0'(z)^{3/2}(\nu)^{-1/2} = B_1 + B_2\sigma + B_3\eta \tag{8}$$

where A_1, A_2, A_3, B_1, B_2, and B_3 are polynomials in M and N.

Equations (7) and (8) are the resulting simultaneous equations which allow the evaluation of the general problem of boundary layer control by distributed surface suction.

The special case for which this method was applied is the determination of the suction velocity distribution over the surface for a prescribed boundary layer thickness variation. By equating the right hand sides of equations (7) and (8) one obtains the following double-parameter equation

$$\phi_1 + \phi_2\sigma + \phi_3\eta = 0 \tag{9}$$

In this equation σ and η are functions of s, prescribed by the physical problem and ϕ_1, ϕ_2 and ϕ_3 are polynomials in M and N. When constant boundary layer thickness is to be maintained over the surface, η is zero, and equation (9) becomes

$$-\sigma = \phi_1/\phi_2 \tag{10}$$

The development of the boundary layer along a solid wall for an arbitrary pressure distribution in the potential flow is another special case of the analysis. Then, v_0 is zero (and, therefore, M is zero), and the momentum and energy equations yield the following two differential equations for the boundary layer thickness variation:

$$Uz' = C_1 + C_2\sigma \tag{11}$$

$$Uz' = C_3 + C_4\sigma \tag{12}$$

where C_1, C_2, C_3, and C_4 are functions of N. These two equations yield

$$-\sigma = \theta_1/\theta_2 \tag{13}$$

and θ_1 and θ_2 are polynomials in N.

Equation (9) is the sum of equations (7) and (8). After the suction velocity distribution is calculated from equation (9) (or equation (10)), it has to be shown that this satisfies not only the sum of the two equations, but also the equations themselves. This may be done by calculating from either equation (7) or (8) the tangents of the v_0 curve at several points along the curve and checking the agreement between the curve and its tangents. For one numerical example worked with the method, excellent agreement has been obtained. A similar procedure has to be followed, if the method is applied to the analysis of boundary layers along solid walls.

For the evaluation of the present analysis, numerical work should be carried out, particularly for cases for which experimental results are available. Unfortunately, most of these experiments are of a classified nature and were not available to the author. The analysis will be published in its complete form when it can be checked with experimental results.

REFERENCE

1. Torda, T. P. Boundary Layer Control by Continuous Surface Suction or Injection. Journal of Mathematics and Physics, October 1952.

University of Illinois

(Received October 3, 1952; Revised August 25, 1953)

A NOTE ON THE PAPER "ON A CLASS OF ITERATIVE PROCEDURES FOR SOLVING NORMAL SYSTEMS OF ORDINARY DIFFERENTIAL EQUATIONS"

By J. Warga

It was brought to the author's attention that the iterative procedure designated by him as "germay's method (1)" was first employed by E. Cotton in 1911 in the *Annales de l'Ecole Normale Superieure*, [3], vol. 28, pp. 473–521. It was subsequently employed by R. Bellman in the *Transactions of the Amer. Math. So.*, vol. 62, no. 3, pp. 359–365, in 1947 and by R. H. Germay in the *Bulletin de la Societé Royale des Sciences de Liègé*, [6], pp. 3–8, in 1949.